CRA 8-5002-FR, Vol 2, Pt. A

TECHNOLOGY UTILIZATION OFFICE
GEORGE C. MARSHALL SPACE FLIGHT CENTER
HUNTSVILLE, ALABAMA

5002

STUDY OF THE DYNAMIC STABILITY AND CONTROL OF LARGE NON-RIGID AEROBALLISTIC VEHICLES

VOLUME II

CRA

CONTROL RESEARCH ASSOCIATES
608 WASHINGTON AVENUE
LAUREL, MARYLAND

PREFACE

This is the second volume of our Final Report for work done under NASA Contract Number NAS 8-5002.

This volume, like the first one, is divided into two parts. Part A is devoted to an analytical study of the time-optimal control problem and is a direct continuation of Part A of the first volume. It consists of results obtained by Daniel C. Lewis and Pinchase Mendelson and was also written by them. Part A of the present volume starts with Chapter 13. The preceding twelve chapters of Part A are contained in the first volume. Bearing this fact in mind, the reader should have no difficulty in following all references which appear in the text.

Part B is devoted to work carried out by J. Gilchrist, J. Schlessinger, G. Campbell and K. A. Ivey on computer simulation of time-optimal control laws. The final editing of this part was carried out by Keith A. Ivey.

PART A

MATHEMATICAL THEORY OF THE TIME OPTIMAL CONTROL PROBLEM

By

DANIEL C. LEWIS

PINCHAS MENDELSON

TABLE OF CONTENTS

PART A

CHAPTER 20 A THREE DIMENSIONAL EXAMPLE OF BANG-BANG CONTROL WITH PHASE COORDINATE CONSTRAINTS

APPENDIX TO PART A

Parametric Equations of the Three-Dimensional Sθitching Manifold for the Fourth Order Linear System with Eigenvalues $(0, 0, -\lambda, \lambda)$.

CHAPTER 13

A CLOSED CONTROL LAW FOR THE SYSTEM $\ddot{x} = \epsilon$, $\epsilon = \pm 1$

1. Derivation of A Closed Control Law For The System $\ddot{x} = \epsilon$

The trivial system

$$\dot{x}_1 = \epsilon, \quad \epsilon = \pm 1, \tag{1}$$

is controlled time-optimally by the function

$$\epsilon = \operatorname{sgn} \sigma, \quad \sigma = -x_1. \tag{2}$$

We shall write $\epsilon_1 = \operatorname{sgn}(-x_1)$.

The function ϵ_1 may be used to define a closed form control law for the second order system.

$$\dot{x}_1 = \epsilon, \quad \dot{x}_2 = x_1, \quad \epsilon = \pm 1 \tag{3}$$

In fact, let

$$\sigma_1 = \epsilon_1$$

$$\sigma_2 = x_2 - \frac{1}{2}\sigma_1 x_1^2 = x_2 + \frac{1}{2}(\operatorname{sgn} x_1)x_1^2 \tag{4}$$

and define

$$\sigma = -\sigma_2 = -[x_2 + \tfrac{1}{2}(\operatorname{sgn} x_1)x_1^2]. \tag{5}$$

Then

$$\epsilon_2 = \mathrm{sgn}\ \sigma \tag{6}$$

is the time-optimal control law for system (3). The fact that (6) does indeed define the time-optimal control law is clear by direct inspection using the known switching curve for system (3). However, certain remarks concerning the definition of σ_2 and σ are in order. Regarding the former we note that

$$\sigma_2 = \begin{cases} y_2 & \text{if } \sigma_1 = 1 \\ -z_2 & \text{if } \sigma_1 = -1 \end{cases} \tag{7}$$

where

$$\left.\begin{aligned} y_1 &= x_1 \\ y_2 &= x_2 - \tfrac{1}{2} x_1^2 \end{aligned}\ \right|\ \begin{aligned} z_1 &= x_1 \\ z_2 &= -(x_2 + \tfrac{1}{2} x_1^2) \end{aligned} \tag{8}$$

are the auxiliary variables defined extensively in previous chapters (cf. Chapter 2, pp. 25-26). On the other hand, equation (5) arises naturally from the known equation of the switching curve of system

(3). The switching curve is composed of 2 branches (leaves) given, respectively, by the equations and inequalities:

$$f(y_2) = y_2 = 0, \quad y_1 < 0$$
$$f(z_2) = z_2 = 0, \quad z_1 < 0$$

The function σ may thus be defined by:

$$\sigma = -f(\sigma_2) = \dot{\sigma}_2 = -(\operatorname{sgn} \sigma_2)\ |\sigma_2|.$$

This procedure can be carried still further and the function ϵ_2 may be used to define the closed form control law for the third order system.

$$\dot{x}_1 = \epsilon, \quad \dot{x}_2 = x_1, \quad \dot{x}_3 = x_2, \quad \epsilon = \pm 1 \qquad (9)$$

We recall first that the auxiliary variables for system (9) are given by [Chapter 2, pp. 25-26]:

$$\begin{aligned} y_1 &= x_1 & z_1 &= -x_1 \\ y_2 &= x_2 - \tfrac{1}{2} x_1^2 & z_2 &= -(x_2 + \tfrac{1}{2} x_1^2) \\ y_3 &= x_3 - x_2 x_1 + \tfrac{1}{3} x_1^3 & z_3 &= -(x_3 + x_2 x_1 + \tfrac{1}{3} x_1^3) \end{aligned} \qquad (10)$$

Let

$$h_2(x_1, x_2, x_3, \eta) = x_2 - \frac{1}{2}\eta x_1^2$$

$$h_3(x_1, x_2, x_3, \eta) = x_3 - \eta x_2 x_1 + \frac{1}{3} x_1^3 \qquad (11)$$

then clearly

$$h_i(x_1, x_2, x_3, +1) = y_i, \qquad i = 2, 3$$

$$h_i(x_1, x_2, x_3, -1) = -z_i, \qquad i = 2, 3$$

Define

$$\sigma_1 = \epsilon_2$$

$$\sigma_2 = h_2(x_1, x_2, x_3, \sigma_1) = x_2 - \frac{1}{2}\sigma_1 x_1^2$$

$$\sigma_3 = h_3(x_1, x_2, x_3, \sigma_1) = x_3 - \sigma_1 x_2 x_1 + \frac{1}{3} x_1^3,$$

in which case

$$\left.\begin{array}{lll} \sigma_i = y_i & \text{if} & \sigma_1 = 1 \\ \sigma_i = -z_i & \text{if} & \sigma_1 = -1 \end{array}\right. \quad i = 2, 3$$

The two leaves of the 2 dimensional switching surface of system (9) are given by

$$R_{22}\colon\ f(y_2, y_3) = y_3^2 + y_2^3 = 0, \quad \frac{y_3}{y_2} < 0, \quad y_1 < -\frac{y_3}{y_2} \tag{12}$$

$$R_{21}\colon\ f(z_2, z_3) = z_3^2 + z_2^3 = 0, \quad \frac{z_3}{z_2} < 0, \quad z_1 < -\frac{z_3}{z_2}$$

Let

$$\sigma = -[(\operatorname{sgn} \sigma_3)\ |\sigma_3|^2 + \sigma_2^3] \tag{12A}$$

We shall show that σ is a closed time-optimal control law for system (9).

Let $P(x_1, x_2, x_3)$ be an arbitrary point in phase space which does not lie on the switching surface of system (9), and whose projection (x_1, x_2) on the (x_1, x_2)-plane does not lie on the switching curve of system (3). The last condition implies that $\sigma_1(P)$ is either +1 or -1. Assume first that $\sigma_1(P) = +1$. The case $\sigma_1(P) = -1$ will be treated later in an analogous fashion.

The variables (x_1, x_2) may be expressed in terms of (y_1, y_2). Therefore, the function $\sigma_1(x_1, x_2) = \epsilon_2(x_1, x_2)$ may be viewed as a function of (y_1, y_2). In the (y_1, y_2)-plane let Σ^* denote the set of all those points P^* such that $\sigma_1(P^*) = \epsilon_2(P^*) = 1$. The set Σ^* is bounded by the two leaves (i) $y_2 = 0$, $y_1 < 0$, (ii)

$z_2 = -(x_2 + \frac{1}{2} x_1^2) = -(y_2 + y_1^2) = 0, \quad z_1 = -y_1 < 0$ (Figure 1).

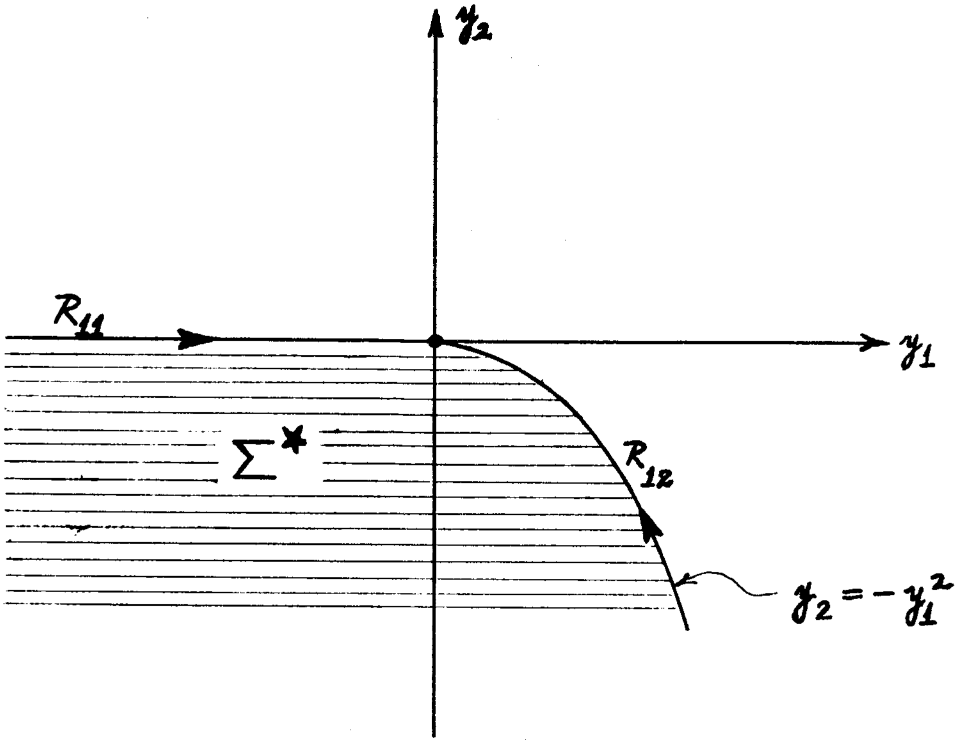


FIGURE 1

In the (y_1, y_2, y_3)-space let Y_3 denote the y_3- axis and let $\Sigma = \Sigma^* \times Y_3$, where $\times$ denotes the Cartesian product. Σ is then the cylindrical set, parallel to the y_3-axis, whose base in the (y_1,y_2)-plane is the set Σ^*. Clearly if $\sigma_1(P) = 1$ then $P \in \Sigma$.

The leaf R_{22} of the two-dimensional switching surface R_2 is bounded by $R_{11} \cup R_{12}$. These one-dimensional leaves may be represented, in terms of the y's, as follows:

$$R_{11}: \quad y_1 < 0, \qquad y_2 = 0, \qquad y_3 = 0, \tag{13}$$

$$R_{12}: \quad \frac{y_3}{y_2} < 0, \quad y_3 + y_2 y_1 = 0, \quad y_3^2 + y_2^3 = 0 \tag{14}$$

(Chapter 2, p. 27 and p. 30.). The projection of R_{11} on the (y_1,y_2)-plane is the negative half of the y_1- axis. To find the projection of R_{12}, solve for y_3 in the first equation in (14) and substitute in the second. The result is $y_2 + y_1^2 = 0$. Moreover, substitution of y_3 into the inequality in (14) yields $y_1 > 0$. Therefore, the projection of $R_{11} \cup R_{12}$ into the (y_1,y_2)-plane coincides with the boundary of Σ^*. The leaf R_{22} is obtained by solving backwards in time, using $\epsilon = +1$, starting on R_{12}. Therefore, in the (y_1,y_2,y_3)-space the leaf R_{22} is a cylindrical set parallel to the y_1-axis and such that along every generator of this leaf the maximum value of y_1 is attained at a point on R_{12}. R_{12} is thus the upper edge of R_{22} with respect to the y_1-axis. It follows that the projection of R_{22} on the (y_1,y_2)-plane is the set Σ^*.

We have arrived at the following configuration: In the (y_1,y_2)-plane there is a set Σ^* which forms a base of a cylindrical set Σ, parallel to the y_3-axis. Within this cylinder lies a leaf of the switching surface (namely R_{22}) in such a way that it is parallel to the y_1-axis and its projection on the (y_1,y_2)-plane is the set Σ^*. Hence R_{22} separates Σ into two distinct parts.

In a completely analogous fashion we note that the variables (x_1,x_2) may be expressed in terms of (z_1, z_2) and that therefore the function $\sigma_1(x_1,x_2)$ may be viewed as a function of (z_1,z_2). Define the set Σ^{**} in the (z_1,z_2)-plane as the set of all points P^* such that $\sigma_1(P^*) = -1$. Let Z_3 be the z_3-axis in the (z_1,z_2,z_3)-space and let $\Sigma' = \Sigma^{**} \times Z_3$. Then R_{21} lies within Σ' and separates it into two distinct parts.

If we map the set Σ' into the (y_1,y_2,y_3)-space via the transformation which relates the y's to the z's (Chapter 2, p. 27) we obtain a set which must be contained in the complement of the set Σ. This may easily be seen as follows: Let P be a point in Σ' with coordinates (z_1^o, z_2^o, z_3^o). Let (x_1^o, x_2^o, x_3^o) be the x-coordinates corresponding to (z_1^o, z_2^o, z_3^o). Then $\sigma_1(x_1^o, x_2^o) = -1$, by definition of Σ'. If (y_1^o, y_2^o, y_3^o) are the y-coordinates of (x_1^o, x_2^o, x_3^o) then clearly $(y_1^o, y_2^o) \epsilon \Sigma^{*c}$, by definition of Σ^*. Hence $(y_1^o, y_2^o, y_3^o) \notin \Sigma$, or equivalently, $(y_1^o, y_2^o, y_3^o) \epsilon \Sigma^c$. It follows, in particular, that R_{21}, when imbedded in the (y_1,y_2,y_3)-space, lies in Σ^c. We conclude that Σ contains exactly one leaf (namely R_{22}) of the switching surface, is divided by it into two parts and does not intersect the second leaf.

As stated above, the leaf R_{22} is parallel to the y_1-axis. Its projection on the (y_2,y_3)-plane satisfies the equation $y_3^2 + y_2^3 = 0$. But $R_{22} \subset \Sigma$ whence $y_2 < 0$ for all points on R_{22}. Thus, by (12), $y_3 > 0$ on R_{22}. Hence the projection of R_{22} on the (y_2,y_3)-plane is given by

$$y_3^2 + y_2^3 = 0$$

$$y_2 < 0, \quad y_3 > 0$$

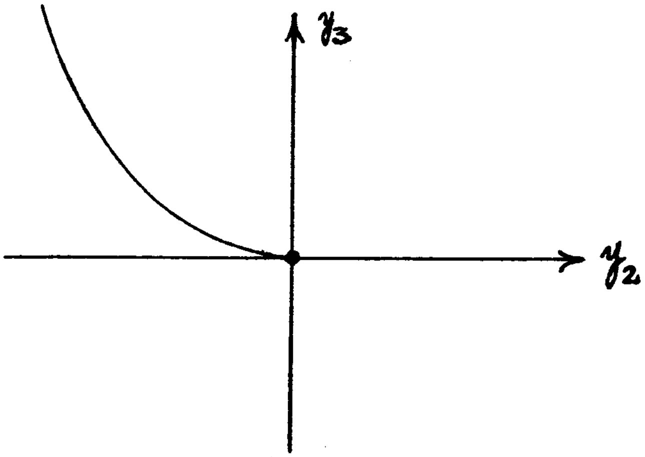


FIGURE 2

We are finally in a position to test the validity of the control law given by (12A). If $\sigma_1(P) = + 1$, then $P \epsilon \Sigma$ and

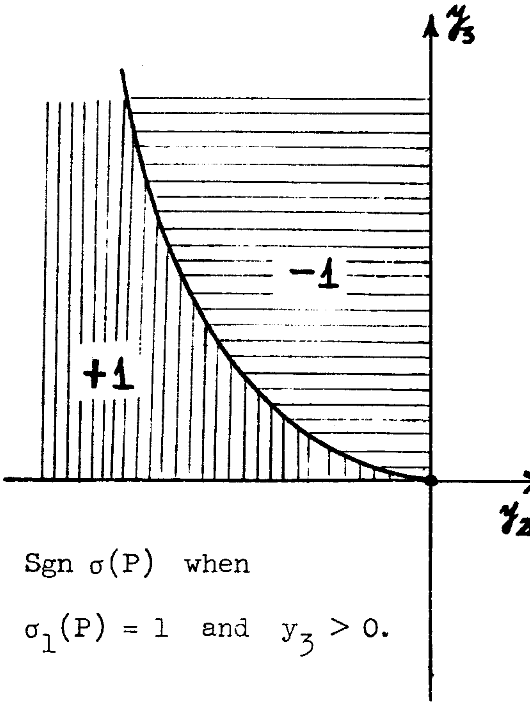
$\sigma(P) = -[(\operatorname{sgn} y_3)|y_3|^2 + y_2^3]$.

There are two possibilities:

(i) $y_3 > 0$, in which case

$$\sigma(P) = -(y_3^2 + y_2^3)$$

and $\operatorname{sgn} \sigma(P)$ is as indicated in Figure 3.



Sgn $\sigma(P)$ when

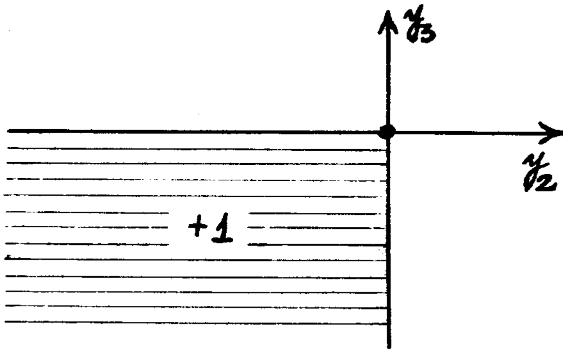
$\sigma_1(P) = 1$ and $y_3 > 0$.

FIGURE 3

(ii) $y_3 < 0$ in which case

$$\sigma(P) = -(-|y_3|^2 - |y_2|^3) > 0$$

and $\operatorname{sgn} \sigma(P) \equiv + 1$. (Figure 4)



Sgn $\sigma(P)$ when

$\sigma_1(P) = 1$ and $y_3 < 0$

FIGURE 4

Thus $\operatorname{sgn} \sigma(P)$ assigns the value $+ 1$ to all points of Σ lying on the one side of R_{22} and the value -1 to all points of Σ

lying on the opposite side. The fact that this particular assignment is the correct one (and not the reverse assignment) follows from a direct inspection of the switching surface (Chapter 11, Plates I, II, III following p.127).

If $\sigma_1(P) = -1$ then $\sigma(P) = -[\{\mathrm{sgn}(-z_3)\}|z_3|^2 + (-z_2)^3]$ and again $\mathrm{sgn}\ \sigma(P)$ assigns the value $+1$ and -1 on opposite sides of R_{21} (in Σ'). Note, however, that in this case

$$\sigma(P) = [(\mathrm{sgn}\ z_3)|z_3|^2 + z_2^3]$$

assigns the opposite values (in the (z_2, z_3)-plane) from those assigned in the (y_2, y_3)-plane in the comparable regions (Figure 5).

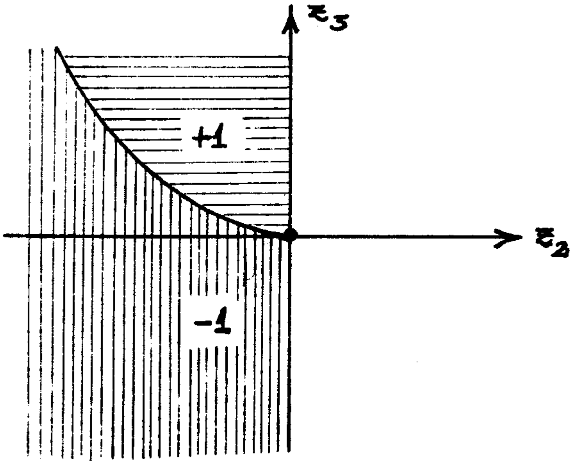


FIGURE 5

This, of course, conforms to the fact that these two leaves correspond to opposite values of control.

2. Appendix: Alternative Proof For The Control Law of § 1.

The proof of the control law for the system $\dot{x}_1 = \epsilon$, $\dot{x}_2 = x_1$, $\dot{x}_3 = x_2$, given in § 1, supersedes the proof given here for the same law. Despite the fact that the proof given here is inferior to the one given in § 1 from the point of view of brevity and elegance, the present proof contains some ideas not occurring in the other proof which might prove to be valuable in more complicated cases.

The determination of ϵ is made as follows:

Let

$$\sigma_1 = \begin{cases} x_2 & \text{if } |x_2| > \frac{1}{2} x_1^2 \\ x_1 & \text{if } |x_2| \leq \frac{1}{2} x_1^2 \end{cases} \tag{1}$$

and let

$$\sigma_3 = x_3 + (\operatorname{sgn} \sigma_1) x_1 x_2 + \frac{1}{2} x_1^3 \tag{2}$$

$$\sigma_2 = x_2 + \frac{1}{2}(\operatorname{sgn} \sigma_1) x_1^2 \tag{3}$$

$$\sigma = -((\operatorname{sgn} \sigma_3)|\sigma_3|^2 + (\operatorname{sgn} \sigma_2)|\sigma_2|^3). \tag{4}$$

Note that $(\operatorname{sgn} \sigma_2)|\sigma_2|^3 = \sigma_2^3$. Hence in this case

$$\sigma = -((\operatorname{sgn} \sigma_3)|\sigma_3|^2 + \sigma_2^3).$$

Finally, we let

$$\epsilon = \text{sgn } \sigma. \tag{5}$$

If the law is correct, the closure of the complete control surface R_2 should be given by the equation $\sigma = 0$. From our previous developments, we know that $R_2 = R_{21} \cup R_{22}$, where R_{21} is characterized by

$$z_1 < - z_3/z_2 \tag{6a}$$

$$z_3/z_2 < 0 \tag{6b}$$

$$z_2^3 + z_3^2 = 0, \tag{6c}$$

where $z_1 = - x_1$, $z_2 = - x_2 - \frac{1}{2} x_1^2$, $z_3 = - x_3 - x_1 x_2 - \frac{1}{3} x_1^3$; and R_{22} is characterized by

$$y_1 < - y_3/y_2 \tag{7a}$$

$$y_3/y_2 < 0 \tag{7b}$$

$$y_2^3 + y_3^2 = 0 \tag{7c}$$

where $y_1 = x_1$, $y_2 = x_2 - \frac{1}{2} x_1^2$, $y_3 = x_3 - x_1 x_2 + \frac{1}{3} x_1^3$.

The first part of our proof is to show that $\sigma = 0$ does indeed yield the closure of R_2. To do this we consider the following four regions of phase space:

$S_1 : \ |x_2| < \frac{1}{2} x_1^2, \qquad x_1 > 0.$

$S_2 : \ x_2 > \frac{1}{2} x_1^2,$

$S_3 : \ |x_2| < \frac{1}{2} x_1^2, \qquad x_1 < 0$

$S_4 : \ x_2 < - \frac{1}{2} x_1^2$

All points of the phase space lie in the closure of the union of these four regions, and no two of these four regions have a common point.

LEMMA 1. If $(x_1,x_2,x_3) \in S_1 \cup S_2$, then $\sigma_1 > 0$, $\sigma_2 = x_2 + \frac{1}{2} x_1^2 = - z_2 > 0$, $\sigma_3 = x_3 + x_1 x_2 + \frac{1}{3} x_1^3 = - z_3$. If, in addition, $\sigma = 0$ for this point, then $-\sigma = (\text{sgn } \sigma_3)\sigma_3^2 + \sigma_2^3 = 0$, $\sigma_3 < 0$, $-\sigma_3^2 + \sigma_2^3 = 0$, which is equivalent to $z_3^2 + z_2^3 = 0$.

The proof of this lemma is omitted, since it is a routine job to check in succession the statements of the lemma in the two cases when $(x_1,x_2,x_3) \in S_1$ and where $(x_1,x_2,x_3) \in S_2$. To do this, we, of course, use the definition, given by (1), (2), (3) and (4) of the various σ's as well as the equations defining z_2 and z_3.

LEMMA 2. If $(x_1,x_2,x_3) \in S_3 \cup S_4$, then $\sigma_1 < 0$, $\sigma_2 = x_2 - \frac{1}{2} x^2 = y_2 < 0$, $\sigma_3 = x_3 - x_1 x_2 + \frac{1}{3} x_1^3 = y_3$. If, in addition, $\sigma = 0$ for this point, then $-\sigma = (\mathrm{sgn}\ \sigma_3)\sigma_3^2 + \sigma_2^3 = 0$, $\sigma_3 > 0$, $+ \sigma_3^2 + \sigma_2^3 = 0$, which is equivalent to $y_3^2 + y_2^3 = 0$.

The proof of this lemma is omitted for reasons analogous to those given for omitting the proof of Lemma 1.

LEMMA 3. If $(x_1,x_2,x_3) \in S_1 \cup S_2$ and if $\sigma = 0$, then (6b) and (6c) are both satisfied.

PROOF. The fact that (6c) is satisfied is clear from the last statement in lemma 1. It is also clear from lemma 1 that $z_3/z_2 = (-\sigma_3)/(-\sigma_2) = \sigma_3/\sigma_2 < 0$, since according to lemma 1, $\sigma_2 > 0$ in $S_1 \cup S_2$ and $\sigma_3 < 0$ if, in addition, $\sigma = 0$.

LEMMA 4. If $(x_1,x_2,x_3) \in S_3 \cup S_4$ and if $\sigma = 0$, then (7b) and (7c) are both satisfied.

PROOF. The fact that (7c) is satisfied is clear from the last statement in lemma 2. It is also clear from lemma 2 that $y_3/y_2 = \sigma_3/\sigma_2 < 0$, since according to lemma 2, $\sigma_2 < 0$ in $S_3 \cup S_4$ while $\sigma_3 > 0$ if, in addition, $\sigma = 0$.

LEMMA 5. If $(x_1,x_2,x_3) \in S_1 \cup S_2$ and if $\sigma = 0$, then (6a) is satisfied.

PROOF. If $(x_1,x_2,x_3) \in S_1$, then, from (1), $\sigma_1 = x_1 = -z_1 > 0$. From lemma 3 and (6b), we have $z_3/z_2 < 0 < -z_1$, which implies (6a).

If $(x_1,x_2,x_3) \in S_2$, then, from lemma 1, we have $z_3 = -\sigma_3$, $z_2 = -\sigma_2$, $\sigma_2 > 0$ and

$$z_1 = - x_1 \tag{8}$$

by definition of z_1. Hence $z_3/z_2 = \sigma_3/\sigma_2$, while we also have by lemma 1, if $\sigma = 0$, $\sigma_3 = -\sigma_2^{3/2}$. Hence

$$z_3/z_2 = -\sigma_2^{1/2} . \tag{9}$$

Again from lemma 1, $\sigma_2 = x_2 + \frac{1}{2} x_1^2 = (x_2 - \frac{1}{2} x_1^2) + x_1^2$. Hence, $\sigma_2 - x_1^2 = x_2 - \frac{1}{2} x_1^2 > 0$ because $(x_1,x_2,x_3) \in S_2$. Therefore, $\sigma_2 > x_1^2$. Therefore, $\sigma_2^{1/2} > |x_1|$, and $-\sigma_2^{1/2} < - |x_1|$. Hence from (8) and (9) we find that

$$\frac{z_3}{z_2} + z_1 = - \sigma_2^{1/2} - x_1 < - |x_1| - x_1 = \begin{cases} -x_1 - x_1 = -2x_1 < 0 & \text{if } x_1 > 0 \\ x_1 - x_1 = 0, & \text{if } x_1 < 0. \end{cases}$$

Thus (6a) holds for any $(x_1,x_2,x_3) \in S_2$ for which $\sigma = 0$. This completes the proof of lemma 5.

LEMMA 6. If $(x_1,x_2,x_3) \in S_3 \cup S_4$ and if $\sigma = 0$, then (7a) is satisfied.

PROOF. If $(x_1,x_2,x_3) \in S_3$, then, from (1), $\sigma_1 = x_1 = y_1 < 0$. From lemma 4 and (7b), we then have $y_3/y_2 < 0 < -y_1$, which implies (7a).

If $(x_1,x_2,x_3) \in S_4$, then, from lemma 2, we have $y_3 = \sigma_3$, $y_2 = \sigma_2$, $\sigma_2 < 0$; and

$$y_1 = x_1 \tag{10}$$

by definition of y_1. Hence $y_3/y_2 = \sigma_3/\sigma_2$, while we also have, by lemma 2, if $\sigma = 0$, $\sigma_3 = (-\sigma_2)^{3/2}$. Hence

$$y_3/y_2 = \frac{(-\sigma_2)^{3/2}}{\sigma_2} = -(-\sigma_2)^{1/2} \tag{11}$$

Again from lemma 2, $\sigma_2 = x_2 - \frac{1}{2}x_1^2 = (x_2 + \frac{1}{2}x_1^2) - x_1^2$. Hence $\sigma_2 + x_1^2 = x_2 + \frac{1}{2}x_1^2 < 0$ since $(x_1,x_2,x_3) \in S_4$. Therefore, $(-\sigma_2)^{1/2} > |x_1|$

and $-(-\sigma_2)^{1/2} < -|x_1|$. Hence, from (10) and (11) we find that

$$\frac{y_3}{y_2} + y_1 = -(-\sigma_2)^{1/2} + x_1 < -|x_1| + x_1 \leq 0.$$

Thus (7a) holds for any $(x_1,x_2,x_3) \in S_4$ for which $\sigma = 0$. This completes the proof of lemma 6.

LEMMA 7. $\sigma_1,\sigma_2,\sigma_3$ are continuous functions of x_1 and x_2 in $S_1 \cup S_3$.

PROOF. From the definitions of S_1 and S_3, we see from (1) that $\sigma_1 = x_1$ in $\overline{S_1 \cup S_3}$. Hence, from (2) and (3), we have

$$\sigma_2 = x_2 + \frac{1}{2}|x_1|x_1 \quad \text{and} \quad \sigma_3 = x_3 + |x_1|x_2 + \frac{1}{3}x_1^3 ,$$

whence the lemma is obvious.

LEMMA 8. If P is on the boundary of any of the regions $S_i (i = 1,2,3,4)$ and if $\sigma = 0$ at P, then every neighborhood of P contains at least one point Q interior to either S_1 or S_3 such that $\sigma = 0$ also at Q.

PROOF. From the definition of regions S_i it is clear that any boundary point of S_2 and S_4 is also a boundary point of S_1 or S_3. Hence we may restrict attention to the case where P is a boundary point of S_1 or S_3.

The regions S_i are cylindrical regions, since the defining inequalities are independent of x_3. Their bases in the x_1, x_2-plane are bounded by the parabolas $x_2 = \pm \frac{1}{2} x_1^2$. The projection P^* of P on this plane must therefore lie on one of these parabolas and the projection of a neighborhood of P must be a small region, containing a set Σ^* of points Q^* lying interior to the bases of either S_1 or S_3 (or possibly both) and having P^* as a limit point(of Σ^*).

Now the function σ is seen from (1), (2), (3) and (4) to be a quadratic polynomial $F(x_3)$ in x_3 with coefficients which are functions of x_1 and x_2. Moreover, these coefficients are continuous in (x_1, x_2) in $S_1 \cup S_3$ by lemma 7. The leading coefficient is $-\text{sgn}\ \sigma_3$.

Suppose first that $\sigma_3(P^*) = 0$. Then $\sigma_3(P^*) = \text{sgn}\ \sigma_3(P^*)x_1x_2 + \frac{1}{3} x_1^3 = 0$. Hence either $x_1 = x_1(P^*) = 0$ or else $\text{sgn}\ \sigma_3(P^*)x_2(P^*) + \frac{1}{3} x_1(P^*)^2 = 0$, whence either $x_2(P^*) = -\frac{1}{3} x_1(P^*)^2$ or $x_2(P^*) = \frac{1}{3} x_1(P^*)^2$. But at $P^* x_2 = \pm \frac{1}{2} x_1^2$. Thus $x_1(P^*) = 0$. Thus we have shown that if $\sigma_3(P^*) = 0$, then $x_1(P^*) = 0$. The only point on

$x_2 = \pm \frac{1}{2} x_1^2$ which satisfies $x_1 = 0$ is the origin. Hence, if $\sigma_3(P^*) = 0$ then P^* is the origin. But then $P = (0,0,x_3)$ where P is further constrained by the requirement that $\sigma(P) = 0$. We then have $\sigma_1(P) = 0$, $\sigma_2(P) = 0$, $\sigma_3(P) = x_3$. Hence the equation $\sigma(P) = 0$, which can here be written $-(\operatorname{sgn} x_3)x_3^2 + 0 = 0$, implies that $x_3 = 0$. Hence P is at the origin. The statement of the lemma is manifestly true if P is at the origin.

The argument contained in the last part of the above paragraph also shows that, regardless of the value of $\sigma_3(P^*)$, the only point P, whose projection into the x_1,x_2-plane is the origin and which satisfies $\sigma(P) = 0$, is the origin itself.

Suppose next that $\sigma_3(P^*) \neq 0$. Then P^* is not the origin and P^* lies on one of the four branches of the parabolas mentioned above. Since σ_3 is a continuous function of x_1,x_2 in $S_1 \cup S_3$, there exists a neighborhood of P^* in $S_1 \cup S_3$ such that $\sigma_3(Q)$ is bounded away from zero throughout that neighborhood, N. Therefore, the polynomial $F(x_3)$ has coefficients continuous in x_1,x_2 throughout N and leading coefficient bounded away from zero. It follows that the roots x_3 of $F(x_3) = 0$ are continuous functions of x_1,x_2 throughout N. Hence, if Q^* is sufficiently close to P^* and $Q^* \in S_3 \cap \{x_3 = 0\}$ (or $Q^* \in S_1 \cap \{x_3 = 0\}$ as the case may be), there exists a root $x_3(Q^*)$ of $F(x_3, Q^*) = 0$ which lies close to the value of x_3 at P. Let $(Q^*, x_3) = Q$. Clearly $Q \epsilon S_3$ (or S_1 as the case may be) and $\sigma(Q) = 0$. This completes the proof of lemma 8.

THEOREM 1. The set of points where $\sigma = 0$ is included in the set $\bar{R}_2 = \bar{R}_{21} \cup \bar{R}_{22}$.

PROOF. From lemmas 3 and 5, we see that, if $\sigma = 0$ at P and if $P \epsilon S_1 \cup S_2$, then $P \epsilon R_{21} \subset \bar{R}_{21}$. From lemmas 4 and 6, we see that, if $\sigma = 0$ at P and if $P \epsilon S_3 \cup S_4$, then $P \epsilon R_{22} \subset \bar{R}_{22}$. From lemma 8, we see, that if $\sigma = 0$ at P and if P is on the boundary of one of the regions $S_i (i = 1,2,3,4)$, then P is a limit point of points Q in either S_1 or S_3 where $\sigma = 0$. Hence, P is in this case a limit point of R_{21} or R_{22}. That is, $P \epsilon \bar{R}_{21} \cup \bar{R}_{22}$ in all possible cases regarding the location of P. Q.E.D.

LEMMA 9. If $P \epsilon R_{21}$, then $P \epsilon (S_1 \cup S_2)^*$ and $\sigma = 0$ at P. Here we use $(S_1 \cup S_2)^*$ to denote $S_1 \cup S_2$ plus the points on the boundary common to S_1 and S_2 less the x_3-axis. It is also the same as the complement of $\overline{S_3 \cup S_4}$ denoted by $\overline{(S_3 \cup S_4)}^c$.

PROOF. R_{21} is characterized by (6a), (6b), and (6c). From (6c), $z_2 < 0$. Hence by definition of z_2, $x_2 + \frac{1}{2} x_1^2 > 0$. By definition of S_4 we thus see that $P \notin \bar{S}_4$. From (6a) and the fact that $z_2 < 0$, we find that $z_1 z_2 > -z_3$. Suppose $P \in \bar{S}_3$, then $x_1 \leqq 0$ and hence $z_1 \geqq 0$ and $z_1 z_2 \leqq 0$. Therefore, $0 \leqq -z_1 z_2 < z_3$. Hence $z_1^2 z_2^2 < z_3^2 = - z_2^3$ by (6c). Hence $z_1^2 < -z_2$. By definition of the z's this means that $x_1^2 < x_2 + \frac{1}{2} x_1^2$. But this last inequality contradicts the assumption that $P \in \bar{S}_3$. Therefore, $P \in \overline{(S_3 \cup S_4)}^c = (S_1 \cup S_2)^*$.

Now, if $P(x_1, x_2, x_3) \in R_{21}$ and therefore $P \in (S_1 \cup S_2)^*$, we find, as in lemma 1, that $\sigma_1 > 0$, $\sigma_2 = x_2 + \frac{1}{2} x_1^2 = -z_2$, $\sigma_2 > 0$, $\sigma_3 = x_3 + x_1 x_2 + \frac{1}{3} x_1^3 = -z_3$. From (6b) and the fact that $z_2 = -\sigma_2 < 0$, we have $z_3 > 0$ and hence $\sigma_3 < 0$. From (6c) we now have $-\sigma_3^2 + \sigma_2^3 = 0$. Since $\sigma_3 < 0$, this last equation may be written $(\text{sgn } \sigma_3)|\sigma_3|^2 + \sigma_2^3 = 0$. Hence, $\sigma = 0$ at P as required.

LEMMA 10. If $P \in R_{22}$, then $P \in (S_3 \cup S_4)^*$ and $\sigma = 0$ at P.

PROOF. R_{22} is characterized by (7a), (7b), and (7c). From (7c), $y_2 < 0$. Hence, by definition of y_2, $x_2 - \frac{1}{2} x_1^2 < 0$. Hence, by definition of S_2, $P \notin \bar{S}_2$. From (7a) and the fact that $y_2 < 0$, we find that $y_1 y_2 > -y_3$. Suppose $P \in \bar{S}_1$, then $x_1 \geqq 0$ and hence

$y_1 \geqq 0$. Therefore $y_1 y_2 \leqq 0$ and $0 \leqq -y_1 y_2 < y_3$. Hence $y_1^2 y_2^2 < y_3^2 = -y_2^3$ by (7c). Therefore $y_1^2 < -y_2$. When this inequality is written in terms of the x's we find that $x_2 + \frac{1}{2} x_1^2 < 0$, which contradicts the supposition that $P \in \bar{S}_1$. Therefore, $P \in (\bar{S}_1 \cup \bar{S}_2)^c = (S_3 \cup S_4)^*$.

Now, if $P(x_1, x_2, x_3) \in R_{22}$ and therefore $P \in (S_3 \cup S_4)^*$, we find, as in lemma 2, that $\sigma_1 < 0$, $\sigma_2 = x_2 - \frac{1}{2} x_1^2 = y_2 < 0$, $\sigma_3 = x_3 - x_1 x_2 + \frac{1}{3} x_1^3 = y_3$. From (7b) and the fact that $y_2 = \sigma_2 < 0$, we have $y_3 > 0$ and hence $\sigma_3 > 0$. From (7c) we now have $\sigma_3^2 + \sigma_2^3 = 0$. Since $\sigma_3 > 0$, this last equation may be written $(\operatorname{sgn} \sigma_3)|\sigma_3|^2 + \sigma_2^3 = 0$. Hence $\sigma = 0$ at P as required.

THEOREM 2. If $P \in \bar{R}_2$, then $\sigma = 0$ at P.

PROOF. $(P \in R_2) \Rightarrow$ (by lemmas 9 and 10) that $\sigma(P) = 0$ and that $P \in (S_1 \cup S_2)^*$ or $(S_3 \cup S_4)^*$. Suppose $P \in \bar{R}_2 - R_2$. Then P is in the boundary of $\bar{R}_{21}$ or $\bar{R}_{22}$. Hence P must lie on the boundary between S_4 and S_1 or between S_2 and S_3 including the possibility that P might lie on the x_3-axis. But in the last case $(P \in \bar{R}_{21}) \Rightarrow y_2^3 + y_3^2 = 0$. However, $(x_1 = x_2 = 0) \Rightarrow (y_1 = y_2 = 0) \Rightarrow$ $(y_3 = 0) \Rightarrow (x_3 = 0)$, which means that P is at the origin. Hence $\sigma(P) = 0$. Thus, we may assume that P is not on the x_3-axis.

Suppose therefore that P lies on the boundary between S_2 and S_3 but let it be not on the x_3-axis. It is also known, that P is on the boundary of R_{21} or R_{22}, i.e., $P \in R_{11} \cup R_{12}$. From the definition of S_2 and S_3 we have

$$x_1 < 0, \quad x_2 = \frac{1}{2} x_1^2 \quad \text{(for } P \text{ on the common boundary of } S_2 \text{ and } S_3\text{)} \tag{12}$$

Since, $x_1 = y_1 < 0$ it follows that $P \in R_{11}$. Hence $y_2 = y_3 = 0$. But $y_2 = x_2 - \frac{1}{2} x_1^2 = \sigma_2$ and $y_3 = x_3 - x_1 x_2 + \frac{1}{3} x_1^3 = \sigma_3$, since $\sigma_1 = x_1 < 0$. Thus $\sigma_2 = \sigma_3 = 0$, so that $\sigma = 0$.

If P lies on the boundary between S_4 and S_1 a similar proof shows again that $\sigma = 0$, so that the proof of Theorem 2 is complete.

Theorem 1 and 2 may be summarized by the statement that the points where $\sigma = 0$ are precisely the points of $\bar{R}_2 = \bar{R}_{21} \cup \bar{R}_{22}$. If σ were continuous and had a non-vanishing gradient, we could finish the proof of the control law by verifying its validity at just one point where $\sigma \neq 0$. It turns out, however, as follows from (1), (2), (3) and (4), that σ may experience discontinuities at points on the surfaces $x_2 = \pm \frac{1}{2} x_1^2$ where $\sigma \neq 0$. Hence, for a complete proof, it appears necessary and sufficient to verify the control law at four separate points, one in each of the four regions S_1, S_2, S_3, and S_4. For we easily see that σ is continuous in each of these regions separately and has a non-vanishing gradient on R_2.

The program was carried out using the following specific points $P_1(6,0, -281\frac{1}{4}) \epsilon S_1$, $P_2(6,19,43) \epsilon S_2$, $P_3(-6,0,281\frac{1}{4}) \epsilon S_3$ and $P_4(-6,-19,-43) \epsilon S_4$. The control law yielded $\epsilon = +1$ at P_1 and P_4 and $\epsilon = -1$ at P_2 and P_3 which is in agreement with the correct values of ϵ at these points as computed in Chapter 11.

CHAPTER 14

A GENERAL THEORY OF CONTROL FUNCTIONS

On Control Laws For Systems Of Arbitrary Order. Reduction To Canonical Form

The detailed analysis of the preceding chapter was intended not so much to establish a control law for a special third order system, but rather it was intended to serve as a stepping stone to the understanding of general systems of order n (not necessarily with zero eigenvalues). It should be stated at the outset that the general problem is by no means solved. However, it is now definitely reduced to a simpler problem (of lower dimensionality) in which all surfaces appear in canonical form. These facts are elucidated below.

Consider a system S of the form

$$\begin{aligned} \dot{x} &= Ax + a\epsilon \\ \dot{x}_{n+1} &= c \cdot x + dx_{n+1} + \alpha\epsilon, \end{aligned} \tag{1}$$

where $x = (x_1, \ldots, x_n)$ is an n-vector, A is an $n \times n$ matrix, a and c are constant n-vectors, d and α are given scalars and ϵ, the control parameter, may take on the values ± 1. The associated n^{th} order system

$$\dot{x} = Ax + a\epsilon \qquad (2)$$

is denoted by S^*. We assume, of course, that system (1) is controllable in some neighborhood of the origin.

We now pose the following general problem: For a given system S suppose that a closed form control law, $\epsilon_n(x)$, for the associated system S^* is completely known. Can $\epsilon_n(x)$ be used to generate a closed-form control law, $\epsilon_{n+1}(x,x_{n+1})$ for the higher order system S ? It is this problem which has now been reduced to a simpler form.

Let us recall the main features of the affirmative solution to this problem as given in Chapter 13 for the special case of a system S of order three with three zero eigenvalues. The vector x in the present formulation corresponds to the vector (x_1,x_2) in Chapter 13, while the vector (x,x_{n+1}) corresponds to (x_1,x_2,x_3). The associated system S^* was given by

$$\dot{x}_1 = \epsilon, \quad \dot{x}_2 = x_1 \qquad (3)$$

and its control law, $\epsilon_2(x)$, was known. The function $\epsilon_2(x)$ was used to define a set of two new functions $\sigma_2(x)$, $\sigma_3(x)$ such that

$$\sigma_i(x) = y_i \quad \text{if} \quad \epsilon_2(x) = +1, \qquad i = 2,3$$

$$\sigma_i(x) = -z_i \quad \text{if} \quad \epsilon_2(x) = -1, \qquad i = 2,3.$$

The vector (x_1,x_2,x_3) may be expressed in terms of (y_1,y_2,y_3) in such a way that x_1 and x_2 are functions of (y_1,y_2) while x_3 is a function of (y_1,y_2,y_3). The same statement holds true if we interchange x_i with y_i, $i = 1,2,3$. In particular, $\epsilon_2(x)$ may be viewed as a function of (y_1,y_2). In the (y_1,y_2)-plane we defined the set Σ^* as the set of all points for which $\epsilon_2 = +1$. We used the set Σ^* as the base of the cylinder Σ in the (y_1,y_2,y_3)-space, where Σ was the Cartesian product $\Sigma^* \times Y_3$. We then showed that Σ contained exactly one leaf of the switching surface of the system S and did not even intersect the other leaf. Finally we showed that the leaf which was contained in Σ had the following two essential properties: (i) it was parallel to the y_1-axis, i.e., it was orthogonal to the (y_2,y_3)-plane, and (ii) it separated Σ into two distinct parts. In complete analogy with the above we also constructed the cylinder Σ' in the (z_1,z_2,z_3)-space and it turned out that Σ' contained the second leaf of the switching surface and did not intersect the first. Moreover, the leaf contained in Σ' had the same properties in the (z_1,z_2,z_3)-space as listed above, namely: (i) it was orthogonal to the (z_2,z_3)-plane, and (ii) it separated Σ' into two distinct parts. This construction was finally used to generate a control function for the given system.

The basic construction summarized above applies to $(n + 1)$st-order systems of type (1).

Associated with system (1) are two sets of auxiliary variables, $(y_1, \ldots, y_{n+1})$ and $(z_1, \ldots, z_{n+1})$. These auxiliary variables reduce the equations of the system to canonical form for $\epsilon = +1$ and $\epsilon = -1$, respectively. We recall that $(y_2, \ldots, y_{n+1})$ and $(z_2, \ldots, z_{n+1})$ are defined by means of n time-independent first integrals of the system S. These first integrals contain ϵ as a parameter and the y's and z's are obtained (except for the introduction of a negative sign designed to guarantee that the transformation from the y's to the z's be involutory) by substituting the values $+1$ and -1 for ϵ in these first integrals. Denote the time-independent first integrals by $h_i(x, x_{n+1}; \epsilon)$, $i = 2, \ldots, n + 1$. Due to the particular form of (1) it is clear that we may pick the first $(n-1)$ of these, namely $h_2, \ldots, h_n$, in such a way that they are independent of x_{n+1}. In other words, $h_2(x, \epsilon), \ldots, h_n(x, \epsilon)$ are time-independent first integrals of S^*. Following our standard procedure we now define canonical variables y_i and z_i for the system S by means of the transformations

$$\begin{aligned} y_i &= h_i(x, x_{n+1}; +1), \qquad i = 2, \ldots, n+1, \\ z_i &= -h_i(x, x_{n+1}; -1), \qquad i = 2, \ldots, n+1, \end{aligned} \tag{3}$$

and it is clear from the above remarks that $(y_2, \ldots, y_n)$ and $(z_2, \ldots, z_n)$ are part of a canonical set of variables for the system S^*. At the risk of redundance, we note again for future reference that $(y_2,\ldots, y_n)$ and $(z_2, \ldots, z_n)$ are independent of x_{n+1}.

When the variables y_1 and z_1 are added to (3), [see Chapter 1] the transformations (3) are invertible in a neighborhood of the origin and it is clear that the inverse transformations are of the form

$$x = \varphi(y_1, \ldots, y_n); \qquad x_{n+1} = \varphi_{n+1}(y_1,\ldots, y_{n+1})$$

$$x = \Psi(z_1, \ldots, z_n); \qquad x_{n+1} = \Psi_{n+1}(z_1,\ldots, z_{n+1}) \qquad (4)$$

Define

$$\Sigma^* = \{(y_1,\ldots, y_n) | \epsilon_n(\varphi(y_1,\ldots,y_n)) = +1\}$$

$$\Sigma^{**} = \{(z_1,\ldots, z_n) | \epsilon_n(\Psi(z_1,\ldots,z_n)) = -1\}.$$

Σ^* is simply the set of all those points in the space of $(y_1,\ldots,y_n)$ at which the function ϵ_n takes on the value $+1$. Σ^{**} is described similarly. Let $Y_{n+1}[Z_{n+1}]$ denote the y_{n+1}-axis $[z_{n+1}$-axis$]$ in the space of y's [z's]. Let

$$\Sigma = \Sigma^{*} \times Y_{n+1} \; ; \quad \Sigma' = \Sigma^{**} \times Z_{n+1} \; .$$

The following theorem embodies the first crucial step in the development of our method.

THEOREM 1. Let R_{ni}, $i = 1,2$, be the two leaves of the n-dimensional switching surface of the system S. Consider R_{ni}, $i = 1,2$, as imbedded in the space of y's. If n is even then $\Sigma \supset R_{n2}$, $\Sigma \cap R_{n1} = 0$ and R_{n2} separates Σ into two distinct parts. If n is odd then $\Sigma \supset R_{n1}$, $\Sigma \cap R_{n2} = 0$ and R_{n1} separates Σ into two distinct parts.

A completely analogous theorem holds for the set Σ', namely;

THEOREM 1' Consider R_{ni}, $i = 1,2$, as imbedded in the space of z's. If n is even then $\Sigma' \supset R_{n1}$, $\Sigma \cap R_{n2} = 0$, and R_{n1} separates Σ' into two distinct parts. If n is odd the subscript 1 is simply replaced by 2.

PROOF. R_{11} coincides with the negative half of the y_1-axis. Therefore R_{11} is parallel to the y_1-axis. R_{21} is obtained by solving backwards in time starting on R_{11} and using $\epsilon = -1$. R_{31} is obtained by solving backwards in time, starting on R_{21} and using $\epsilon = +1$. Hence R_{31} is parallel to the y_1-axis. It is clear therefore that the leaf R_{k1} of the k-dimensional switching surface R_k is parallel to the y_1-axis iff k is odd. Similarly it is clear that the leaf R_{k2} of the k-dimensional switching surface R_k is parallel to the y_1-axis iff k is even. In particular, one has that R_{n2} is parallel to the y_1-axis iff n is even and R_{n1} is parallel to the y_1-axis iff n is odd, where $(n + 1)$ is the order of the system S. We may assume, without loss of generality, that n is odd. The treatment of the case when n is even is completely analogous.

Let R_{ki}, R^*_{ki}, $i = 1,2$, denote the leaves of the k-dimensional switching surfaces in the systems S and S^*, respectively. The leaf R_{11} is given by $y_1 < 0$, $y_2 = \dots = y_{n+1} = 0$. The leaf R^*_{11} is given by $y_1 < 0$, $y_2 = \dots = y_n = 0$. It is clear from (4) that in a neighborhood of the origin the (invertible) transformation from $(y_1, \dots, y_{n+1})$ to $(z_1, \dots, z_{n+1})$ is such that $(y_1, \dots, y_n)$ is independent of z_{n+1} while $(z_1, \dots, z_n)$ is independent of y_{n+1}. It follows that if we express the equations of R_{11} in terms of $z = (z_1, \dots, z_n)$ and z_{n+1} we would get

(1) one inequality which does not contain z_{n+1}

(2) (n-1) equalities which do not contain z_{n+1}

(3) one equality which contains z and z_{n+1} .

It is clear from the previous discussion that R^*_{11} , when expressed in terms of z is identical with items (1) and (2) above.

Assuming that $n \geq 3$ we now proceed to eliminate z_1 by using the $(n-1)$ equations of R_{11} which do not contain z_{n+1} . Once this elimination is effected we get a new representation of R_{11} of the following type:

(1) $g_1(z) < 0$, where $z = (z_1, \ldots, z_n)$

(2) $A(z)z_1 + B(z) = 0$

(3) $f_i(z) = 0, \ i = 3, \ldots, n$

(4) $f_{n+1}(z, z_{n+1}) = 0$,

where f_{n+1} is independent of z_1, as are g_1, A, B and f_i, $i = 3, \ldots, n$.

On the other hand, the representation of R^*_{11} reduces to

(1) $g_1(z) < 0$

(2) $A(z)z_1 + B(z) = 0$

(3) $f_i(z) = 0, \ i = 3, \ldots, n$,

where $g_1(z)$, $A(z)$, $B(z)$, $f_i(z)$, $i = 3,\ldots, n$, are identical with those above.

The equations and inequalities of R_{21}, when expressed in terms of the z's are therefore of the form:

(1) $g_i(z) < 0, \quad i = 1,2$

(2) $f_i(z) = 0, \quad i = 3,\ldots,n,$

(3) $f_{n+1}(z, z_{n+1}) = 0,$

while the equations of R^*_{21} are given by the first two items alone.

If we express the equations of R_{21} in terms of the y's we obtain expressions in the form:

(1) $g_i'(y) < 0, \quad i = 1,2, \quad y = (y_1,\ldots,y_n),$

(2) $f_i'(y) = 0, \quad i = 3,\ldots,n,$

(3) $f'_{n+1}(y, y_{n+1}) = 0,$

while the equations of R^*_{21} are given by the first two items alone. If $n \geqq 4$ we now eliminate y_1 from the equations which do not contain y_{n+1} in which case the expressions which define R_{21} are replaced by the appropriate equivalent set in which only the last equation contains y_{n+1}. The expressions defining R^*_{21} are identical with those, except for the absence of the last equation containing y_{n+1}.

Proceeding thus by induction we finally arrive at the leaf $R_{n-2,1}$

which is given by

(1) $g_i'(z) < 0, \quad i = 1, \ldots, n-2.$

(2) $f_{n-1}(z) = 0, \quad f_n(z) = 0$

(3) $f_{n+1}(z, z_{n+1}) = 0,$

while the leaf $R^*_{n-2,1}$ is given by the first two items alone.

We now eliminate z_1 by using the two equations $f_{n-1}(z) = f_n(z) = 0$ to obtain the following equivalent representation of $R_{n-2,1}$:

(1) $g_i(z_2, \ldots, z_n) < 0, \quad i = 1, \ldots, n-2,$

(2) $A(z_2, \ldots, z_n)z_1 + B(z_2, \ldots, z_n) = 0,$

(3) $f(z_2, \ldots, z_n) = 0,$

(4) $k(z_2, \ldots, z_{n+1}) = 0.$

At the same time, the representation of $R^*_{n-2,1}$ consists of the first three items alone.

Thus $R_{n-1,1}$ is given by

(1) $g_i(z_2, \ldots, z_n) < 0, \quad i = 1, \ldots, n-2;$

$$z_1 < -B(z_2, \ldots, z_n)/A(z_2, \ldots, z_n),$$

(2) $f(z_2,\ldots,z_n) = 0$

(3) $k(z_2,\ldots,z_n) = 0$,

while the representation of $R^*_{n-1,1}$ consists of the first two items alone.

Expressing these equations and inequalities in terms of the y's one gets, finally

$$R_{n-1,1}: \quad \begin{array}{ll} (1) & g_i'(y) < 0, \quad i = 1,\ldots,n-1 \\ (2) & f'(y) = 0 \\ (3) & k'(y,y_{n+1}) = 0, \end{array}$$

whereas

$$R^*_{n-1,1}: \quad \begin{array}{ll} (1) & g_i'(y) < 0, \quad i = 1,\ldots,n-1 \\ (2) & f'(y) = 0. \end{array}$$

Since n is odd, the n-dimensional leaf R_{n1} of the system S, when imbedded in the space of (y,y_{n+1}), is parallel to the y_1-axis. It consists of all those points which lie on trajectories which are obtained by moving backwards in time, starting at $R_{n-1,1}$, with $\epsilon = +1$. These trajectories are straight half-lines parallel to the y_1-axis. In other words, R_{n1} is a cylindrical set (parallel to the

y_1-axis) for which $R_{n-1,1}$ forms the upper edge with respect to y_1.

The set Σ^* consists of all those points in the space of y for which $\epsilon_n = +1$. Therefore, if $P^* \epsilon \Sigma^*$ then the optimal trajectory through P^* rises parallel to the y_1-axis until it intersects $R^*_{n-1,1} \cup R^*_{n-1,2}$. Since n is odd, $n-1$ is even whence $R_{n-1,2}$ is parallel to the y_1-axis. The motion through P^* must therefore intersect R^*_{n-1} on $R^*_{n-1,1}$. Conversely, every point which is obtained by moving downward (with respect to y_1) from $R^*_{n-1,1}$, parallel to the y_1-axis, is in Σ^*. It follows that Σ^* is a cylindrical set (parallel to the y_1-axis) for which $R_{n-1,1}$ forms the upper edge with respect to y_1.

It is clear from the last representation of the leaves $R_{n-1,1}$ and $R^*_{n-1,1}$ that $R_{n-1,1} \subset R^*_{n-1,1} \times Y_{n+1}$, where Y_{n+1} represents the y_{n+1}-axis. This fact, when combined with the observations of the two preceding paragraphs yields that R_{n1} is contained in Σ.

One now shows, in a manner analogous to the above, that R_{n2}, when imbedded in the space of (z, z_{n+1}) is contained in the set Σ'.

Let $Q(z^o_1, \ldots, z^o_{n+1}) \epsilon R_{n2}$ and let the x- and y- coordinates of Q be $(x^o_1, \ldots, x^o_{n+1})$ and $(y^o_1, \ldots, y^o_{n+1})$, respectively. Since

$Q \in \Sigma'$ it follows that $\epsilon_n(x_1^o,\ldots,x_n^o) = -1$, whence $(y_1^o,\ldots,y_n^o) \in \Sigma^{*c}$. It follows that $(y_1^o,\ldots,y_{n+1}^o) \in \Sigma^c$. Thus R_{n2}, when imbedded in the space of (y,y_{n+1}) does not intersect Σ.

The n-dimensional switching surface R_n separates the (n+1)-dimensional space of (x,x_{n+1}) into two distinct parts (Chapter 12). This property is preserved when R_n is imbedded in the space of (y,y_{n+1}). But since $\Sigma \cap R_n = R_{n1}$ it follows that R_{n1} separates Σ into two distinct parts. This completes the proof of Theorem 1.

The proof of Theorem 1' is almost identical, except for some obvious modifications. The details are left to the reader.

We recall our general problem: Given a control function $\epsilon_n(x)$ for the system S^*, find a control function $\epsilon_{n+1}(x,x_{n+1})$ for the system S.

Suppose n is even. We note that the equation and inequalities which define R_{n2} in terms of y's and R_{n1} in terms of z are identical except for the interchange of y's and z's and vice versa. Thus the configuration of Σ and R_{n2} in the space of y's is identical with the configuration of Σ' and R_{n1} in the space of z's. A similar statement holds for the case when n is odd.

Suppose again that n is even. The leaf R_{n2} is orthogonal to the hyperplane $y_1 = 0$. A control function _within_ Σ would therefore be independent of y_1. Let $\operatorname{sgn} F(y_2,\ldots,y_{n+1})$ be a control function within Σ. The reader will readily convince himself that, on account of the preceding paragraph, $\operatorname{sgn}[-F(z_2,\ldots,z_{n+1})]$ would then be a control function within Σ'. Let

$$F(\xi_2,\ldots,\xi_{n+1}) = G(-\xi_2,\ldots,-\xi_{n+1}) \qquad (5)$$

then $\operatorname{sgn}[-G(-z_2,\ldots,-z_{n+1})]$ is a control function within Σ'. Let

$$\sigma_1 = \epsilon_n(x)$$

$$\sigma_i = h_i(x,x_{n+1}; \sigma_1(x)), \quad i = 2,\ldots,n+1 \qquad (6)$$

then clearly

$$\sigma_i = y_i \quad \text{whenever} \quad \epsilon_n(x) = +1, \quad i = 2,\ldots,n+1,$$

$$\sigma_i = -z_i \quad \text{whenever} \quad \epsilon_n(x) = -1, \quad i = 2,\ldots,n+1.$$

Define

$$\sigma(x,x_{n+1}) = (1 + \sigma_1)F(\sigma_2,\ldots,\sigma_{n+1}) - (1-\sigma_1)G(\sigma_2,\ldots,\sigma_{n+1}) \qquad (7)$$

then

$$\epsilon_{n+1}(x,x_{n+1}) = \operatorname{sgn} \sigma \qquad (8)$$

is a control function of S. We summarize these results in Theorem 2.

THEOREM 2. Let $\operatorname{sgn} F(y_2,\ldots,y_{n+1})$ be a control function within Σ. Let G be defined as in (5) and let σ_i, $i = 1,\ldots,n+1$, be defined by (6). Then $\epsilon_{n+1} = \operatorname{sgn} \sigma$ is a control function for S, where σ is defined by (7).

The import of Theorems 1 and 2 is to reduce the problem of finding a control function throughout the whole of phase space to that of finding a control function (in terms of the y's) in the set Σ alone. This task is simplified somewhat further by the fact that Σ is a cylindrical set which is parallel to the y_{n+1}-axis and the leaf which separates it (see Theorem 1) is orthogonal to the hyperplane $y_1 = 0$. It is therefore sufficient to consider the projection of this leaf in the hyperplane $y_1 = 0$. Finally, since Σ is a cylindircal set parallel to the y_{n+1}-axis it is sufficient to search for the two sides of the separating leaf in the direction of Y_{n+1} alone. This would certainly tend to simplify greatly the problem of finding a suitable $F(y_2,\ldots,y_{n+1})$. However, a general procedure for obtaining this function is not yet available.

One final remark is in order. The function σ is indeterminate on the set on which $\epsilon_n(x) = 0$. It also vanishes on the switching surface of the system S. However, these two sets have n-dimensional measure zero and will therefore not affect the effectiveness of the function σ in any significant manner.

CHAPTER 15

A CONTROL FUNCTION FOR CONTROLLABLE LINEAR SYSTEMS WITH EIGENVALUES $0,\lambda,-\lambda$

1. Derivation Of A Control Function For The Third Order System With Eigenvalues $0, \lambda, -\lambda$.

In the preceding chapter we developed certain aspects of a general theory of control functions. This theory was first illustrated in Chapter 13 where we obtained a control function for a third order system with three zero eigenvalues. In the present chapter we shall use the same basic approach to obtain a control function for the third order system

$$\begin{aligned} \dot{x}_1 &= \epsilon \\ \dot{x}_2 &= \lambda x_2 + \epsilon, \quad \lambda > 0 \\ \dot{x}_3 &= -\lambda x_3 + \epsilon, \end{aligned} \tag{1}$$

where ϵ, the control parameter, may take on the values ± 1, and λ is real. This system was discussed extensively in Chapter 9 of this Final Progress Report [FPR] and the reader is referred to it for the equations of the switching surface and the definition of the auxiliary variables used below.

Denote the system (1) by S and the system

$$\begin{aligned} \dot{x}_1 &= \epsilon \\ \dot{x}_2 &= \lambda x_2 + \epsilon, \quad \lambda > 0 \end{aligned} \tag{2}$$

by S*. We recall that our procedure requires that before we attempt to define a control function for the system S we first find a control function for the system S*. We shall therefore start by addressing ourselves to this lower order system.

The function

$$\epsilon_1 = \mathrm{sgn}(-x_1) \tag{3}$$

is clearly a time-optimal control function for the system

$$S^{**}: \quad \dot{x}_1 = \epsilon, \qquad \epsilon = \pm 1. \tag{4}$$

We shall use the function ϵ_1 to define a control function for the system S*.

Associated with the system S* are the auxiliary variables (FPR, Vol. 1, Ch. 9):

$$\begin{aligned} y_1 &= x_1 & z_1 &= -x_1 \\ y_2 &= -1 + e^{-\lambda x_1}(\lambda x_2 + 1) & z_2 &= -[1 + e^{\lambda x_1}(\lambda x_2 - 1)] \end{aligned} \tag{5}$$

which reduce S* to canonical form when $\epsilon = +1$ and $\epsilon = -1$, respectively. We shall refer to the transformation from the x's to the y's as (5A). The inverse of (5A) is given by

$$x_1 = y_1$$

$$x_2 = \frac{1}{\lambda} [e^{\lambda y_1}(y_2 + 1) - 1] \tag{6}$$

and the transformation which gives the z's in terms of the y's can easily be found to be

$$z_1 = -y_1$$

$$z_2 = -[1 + e^{\lambda y_1}((y_2 + 1\}e^{\lambda y_1} - 2)] \tag{7}$$

It is easy to see that the controllable region in the (x_1, x_2)-space for system S* does not consist of the whole of phase-space, but only of the strip $|x_2| < \frac{1}{\lambda}$ (Fig. 1). Points lying outside this strip cannot be controlled with $|\epsilon| \leq 1$.

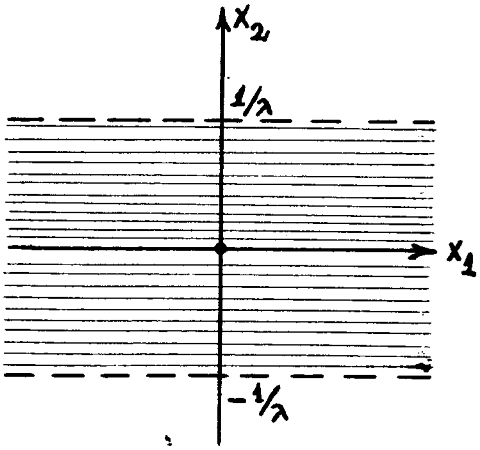


FIGURE 1

The line $x_2 = -\frac{1}{\lambda}$ is mapped by (5A) onto the line $y_2 = -1$. The line $x_2 = \frac{1}{\lambda}$ is mapped by (5A) onto the curve $y_2 = 2e^{-\lambda y_1} -1$. The origin is mapped into the origin. Hence the shaded region of Fig. 1 is mapped onto the shaded region in Fig. 2. This, then, is the controllable region in the (y_1,y_2)-plane.

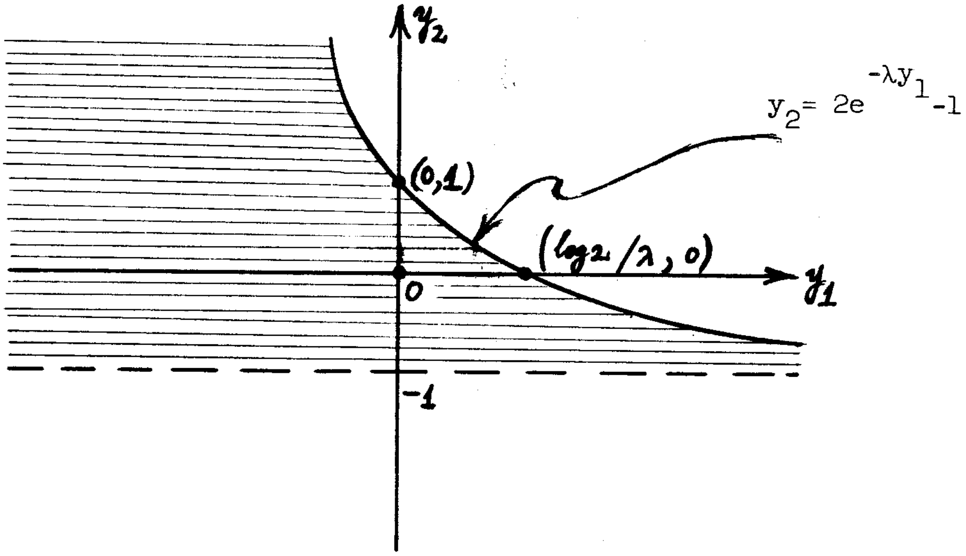


FIGURE 2

The switching curve for system S* is made up of two leaves given, respectively, by

R_{11}: $y_1 < 0, \quad y_2 = 0$

R_{12}: $z_1 < 0, \quad z_2 = 0$

The equations of R_{12}, when written in terms of the y's, take the form

$$R_{12}: \quad y_1 > 0, \quad -[1 + e^{\lambda y_1}(\{y_2 + 1\}e^{\lambda y_1} -2)] = 0$$

or

$$y_1 > 0, \quad e^{2\lambda y_1}(y_2 + 1) - 2e^{\lambda y_1} + 1 = 0 \tag{8}$$

Solving for $e^{\lambda y_1}$ yields

$$e^{\lambda y_1} = \frac{1 \pm \sqrt{-y_2}}{1 + y_2}$$

Clearly y_2 must satisfy $y_2 < 0$. On the other hand we see from Fig. 2 that on R_{12} we must have $y_2 > -1$ if the leaf in question is to lie within the controllable region. Thus $-1 < y_2 < 0$ on R_{12}.

If $-1 < y_2 < 0$ then $|y_2|^{\frac{1}{2}} > |y_2|$ whence

$$\frac{1 - \sqrt{-y_2}}{1 + y_2} = \frac{1 - |y_2|^{\frac{1}{2}}}{1 - |y_2|} < 1$$

However, $y_1 > 0$ on R_{12}, which implies that $e^{\lambda y_1} > 1$ on R_{12}. It follows that the proper leaf is given by

$$R_{12}\colon\ y_1 > 0, \qquad e^{\lambda y_1} = \frac{1 + \sqrt{-y_2}}{1 + y_2} = \frac{1 + |y_2|^{\frac{1}{2}}}{1 - |y_2|} = \frac{1}{1-(-y_2)^{\frac{1}{2}}}$$

or equivalently

$$R_{12}\colon\ y_1 > 0, \quad y_2 = -e^{-2\lambda y_1} + 2e^{-\lambda y_1} - 1 \tag{9}$$

On R_{12} we thus have

$$y_1 > 0; \quad \frac{dy_2}{dy_1} = 2\lambda e^{-\lambda y_1}(e^{-\lambda y_1} -1); \quad \frac{d^2 y_2}{dy_1^2} = 2\lambda^2 e^{-\lambda y_1}(1-2e^{-\lambda y_1}).$$

Hence on R_{12}, y_2 is a monotonically decreasing function, $y_2'(0) = 0$, $y_2'(y_1) \to 0$ as $y_1 \to \infty$; y_2'' is negative in a neighborhood of the origin, positive for y_1 large enough and zero at $y_1 = \frac{\log 2}{\lambda}$. It follows that the leaf R_{12} has the aspect depicted in Fig. 3. The situation in the (z_1, z_2)-plane is identical except for the fact that R_{12} would be replaced by R_{11} and vice versa and that the sign of ϵ in corresponding regions would be reversed.

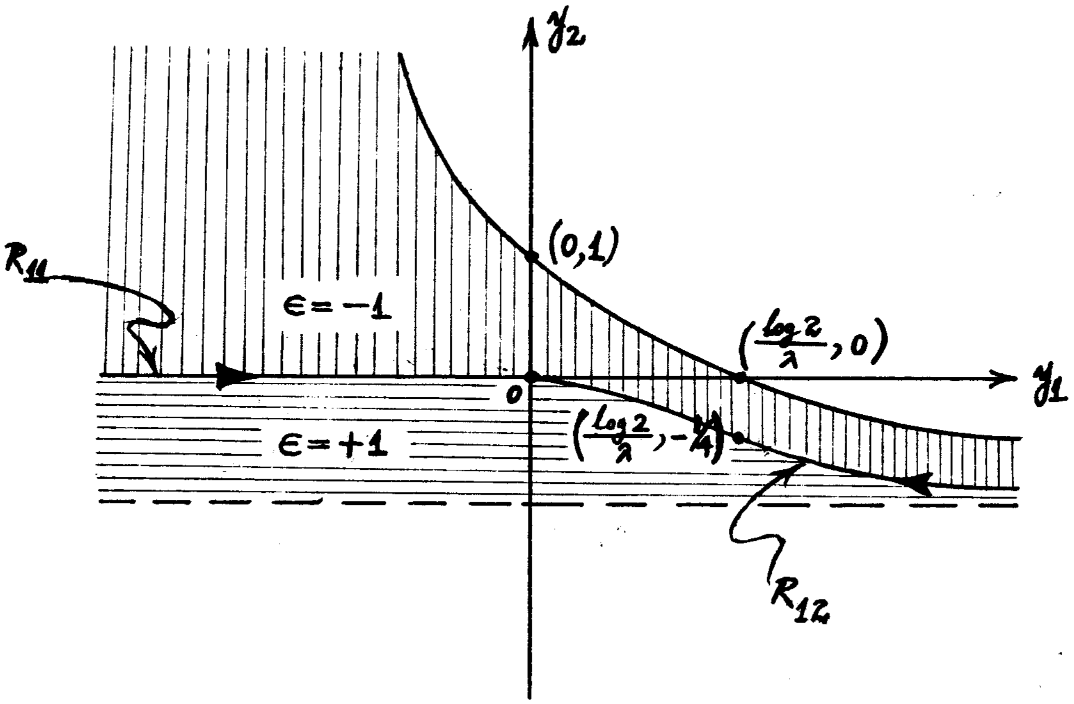


FIGURE 3

Let $P(x_1, x_2)$ be an arbitrary point in phase space. If $x_1 \neq 0$ then either $\epsilon_1(x_1) = 1$ or $\epsilon_1(x_1) = -1$.

Suppose $\epsilon_1 = 1$. Then $x_1 < 0$, whence $y_1 < 0$. It follows from Fig. 3 that $\epsilon = +1$ if $y_2 < 0$ and $\epsilon = -1$ if $y_2 > 0$. Thus

$$\epsilon_1 = 1 \Rightarrow \left\{ \begin{array}{lll} y_2 < 0 & \Rightarrow & \epsilon = +1 \\ y_2 > 0 & \Rightarrow & \epsilon = -1 \end{array} \right\} \Rightarrow \epsilon = \operatorname{sgn}(-y_2) \qquad (10)$$

Suppose $\epsilon_1 = -1$. Then $z_1 < 0$. Hence

$$\epsilon_1 = -1 \Rightarrow \left\{ \begin{array}{lll} z_2 < 0 & \Rightarrow & \epsilon = -1 \\ z_2 < 0 & \Rightarrow & \epsilon = +1 \end{array} \right\} \Rightarrow \epsilon = \operatorname{sgn} z_2 \qquad (11)$$

Let

$$h_2(x_1,x_2,\eta) = -\eta + e^{-\eta\lambda x_1}(\lambda x_2 + \eta) \qquad (12)$$

then clearly

$$h_2(x_1,x_2,+1) = y_2; \quad h_2(x_1,x_2,-1) = -z_2$$

Define

$$\sigma^*(x_1,x_2) = -h_2(x_1,x_2,\epsilon_1) \qquad (13)$$

then clearly

$$\epsilon_2(x_1,x_2) = \operatorname{sgn} \sigma^* \qquad (14)$$

is a time-optimal control function for the system S*.

For a more complete exposition of the phase portrait of the trajectories in the controllable region in the (y_1,y_2)-plane the reader is referred to the Appendix to this chapter. This detailed phase-portrait, while of some interest in itself, is unnecessary for the discussion which follows.

The auxiliary variables for system S are given by (FPR, Vol. 1, p. 107)

$$\begin{array}{l|l} y_1 = x_1 & z_1 = -x_1 \\ y_2 = -1 + e^{-\lambda x_1}(\lambda x_2 + 1) & z_2 = -(1 + e^{\lambda x_1}(\lambda x_2 - 1)) \\ y_3 = 1 + e^{\lambda x_1}(\lambda x_3 - 1) & z_3 = -(-1 + e^{-\lambda x_1}(\lambda x_3 + 1)) \end{array} \tag{15}$$

Let

$h_2(x_1, x_2, x_3; \eta) = -\eta + e^{-\eta\lambda x_1}(\lambda x_2 + \eta) = h_2(x_1, x_2, \eta)$ as given previously by (12), and

$$h_3(x_1, x_2, x_3; \eta) = \eta + e^{\eta\lambda x_1}(\lambda x_3 - \eta) \tag{16}$$

Then clearly

$$h_i(x_1, x_2, x_3; +1) = y_i, \quad i = 2,3$$

$$h_i(x_1, x_2, x_3; -1) = -z_i, \quad i = 2,3 \tag{17}$$

Define

$$\sigma_1(x_1,x_2,x_3) = \epsilon_2(x_1,x_2),$$

$$\sigma_2(x_1,x_2,x_3) = h_2(x_1,x_2,x_3;\ \sigma_1), \tag{18}$$

$$\sigma_3(x_1,x_2,x_3) = h_3(x_1,x_2,x_3;\ \sigma_1)$$

Following the procedure developed in Chapter 14 we note that (x_1,x_2) are functions of (y_1,y_2) and that therefore $\epsilon_2(x_1,x_2)$ can be viewed as a function defined in the (y_1,y_2)-plane. Let Σ^* be that subset in the controllable region in the (y_1,y_2)-plane for which $\epsilon_2 = +1$. The set Σ^* is the horizontally shaded region in Fig. 3. We denote the y_3-axis by Y_3 and define Σ as the Cartesian product $\Sigma^* \times Y_3$. Consider the leaves R_{21} and R_{22} of the two-dimensional switching surface of the system S as imbedded in the (y_1,y_2,y_3)-space. It was proved in Chapter 14 that Σ contains the leaf R_{22}, is divided by it into two distinct parts and does not intersect the other leaf. Furthermore, the leaf R_{22} is parallel to the y_1-axis.

In a completely analogous fashion we note that (x_1,x_2) are functions of (z_1,z_2) and that therefore ϵ_2 is defined in the (z_1,z_2)-plane. Let Σ^{**} be the set of points in the (z_1,z_2)-plane for which $\epsilon_2 = -1$, let Z_3 designate the z_3 -axis and let

$\Sigma' = \Sigma^{**} \times Z_3$. Consider the leaves R_{21} and R_{22} as imbedded in the (z_1,z_2,z_3)-space. Then Σ' contains the leaf R_{21}, is divided by it into two parts and does not intersect the other leaf. Furthermore, R_{21} is parallel to the z_1-axis. Finally we note for future reference that the equations and inequalities which define R_{22} in the space of y's are identical with the equations and inequalities which define R_{21} in the space of z's except for the interchange of y's by z's and vice versa. The last assetion follows from the fact that the transformation from the y's to the z's is involutory and can also be seen directly for this special system by referring to FPR, Vol. 1, pp. 109-110.

The leaf R_{22} is defined in the space of y's by (FPR, Vol. 1, p. 110)

$$R_{22}: \begin{cases} e^{\lambda y_1} < \dfrac{y_2 - y_3 - y_2 y_3}{2y_2} \\ y_2 y_3 + y_2^2 y_3 + y_2^2 < 0 \\ (y_3 - y_2 + y_2 y_3)^2 + 4y_2 y_3 = 0. \end{cases} \qquad (19)$$

If $P(y_1,y_2,y_3) \in R_{22} \subset \Sigma$ then $(y_1,y_2) \in \Sigma^*$ whence $-1 < y_2 < 0$.

Solving the last equation in (19) for y_3 yields two branches, y_3^{I} and y_3^{II}, in the (y_2,y_3)-plane, whose equations are given by

$$y_3^{I} = \frac{y_2^2 - y_2 + 2y_2\sqrt{-y_2}}{(1 + y_2)^2} = \frac{y_2^2 - y_2 - 2(-y_2)^{3/2}}{(1 + y_2)^2}\ , \tag{20}$$

$$y_3^{II} = \frac{y_2^2 - y_2 - 2y_2\sqrt{-y_2}}{(1 + y_2)^2} = \frac{y_2^2 - y_2 + 2(-y_2)^{3/2}}{(1 + y_2)^2} \tag{21}$$

Let $\xi = (-y_2)^{\frac{1}{2}}$. As y_2 ranges from 0 to -1, ξ ranges from 0 to 1. Moreover, $d\xi/dy_2 = -1/2\xi$ and

$$y_3^{I} = \frac{\xi^2}{(1 + \xi)^2}\ ; \qquad y_3^{II} = \frac{\xi^2}{(1 - \xi)^2}$$

Hence

$$\frac{dy_3^{I}}{dy_2} = -\frac{1}{(1 + \xi)^3}\ ; \qquad \frac{dy_3^{II}}{dy_2} = -\frac{1}{(1 - \xi)^3}$$

and

$$\frac{d^2y_3^{I}}{dy_2^2} = -\frac{3}{2\xi(1 + \xi)^4}\ ; \qquad \frac{d^2y_3^{II}}{dy_2^2} = \frac{3}{2\xi(1 - \xi)^4}$$

Thus

$$y_3^I(0) = y_3^{II}(0) = 0; \qquad y_3^I \to \frac{1}{4} \text{ and } y_3^{II} \to +\infty \text{ as } y_2 \to -1$$

$$\left.\frac{dy_3^I}{dy_2}\right|_{y_2=0} = \left.\frac{dy_3^{II}}{dy_2}\right|_{y_2=0} = -1$$

$$\frac{dy_3^I}{dy_2} \to -\frac{1}{8} \text{ as } y_2 \to -1; \qquad \frac{dy_3^{II}}{dy_2} \to -\infty \text{ as } y_2 \to -1$$

$$\frac{d^2y_3^I}{dy_2^2} < 0, \qquad \frac{d^2y_3^{II}}{dy_2^2} > 0 \quad \text{for all } y_2 \epsilon (-1, 0)$$

A geometric representation of y_3^I and y_3^{II} is given in Figure 4.

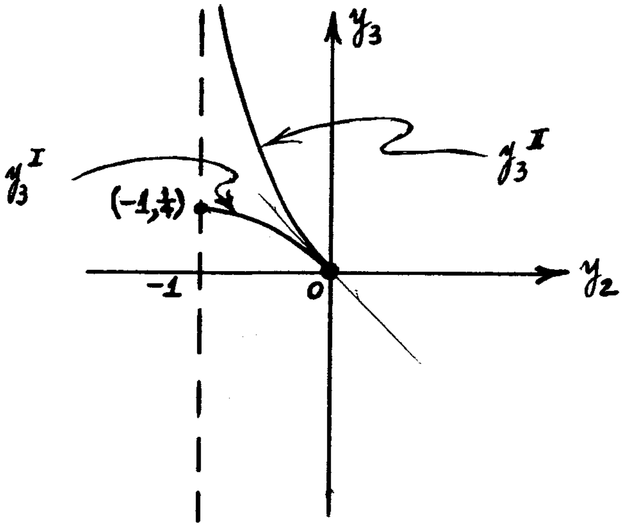


FIGURE 4

The leaf R_{22} satisfies the inequality (see (19))

$$y_2(y_3+ y_2y_3+ y_2) < 0 \tag{22}$$

However, since $y_2 < 0$ on R_{22}, (22) may be replaced by

$$y_3+ y_2y_3+ y_2 = y_3(1 + y_2) + y_2 > 0 \tag{23}$$

On the leaf y_3^I we have

$$y_3^I(1 + y_2) + y_2= \frac{\xi^2}{(1 + \xi)^2} (1-\xi^2)-\xi^2 = - \frac{2\xi^3}{1 + \xi} < 0, \quad \xi \in (0,1)$$

Hence y_3^I is a spurious branch. The proper branch is given by y_3^{II} .

Let $f(y_2,y_3) = [y_3(1 + y_2) -y_2]^2 + 4y_2y_3$

$$= y_3^2(1 + y_2)^2 + 2y_2(1-y_2)y_3+ y_2^2 \tag{24}$$

Then f vanishes on y_3^I and y_3^{II} . Moreover,

$$\frac{\partial f}{\partial y_3} = 2y_3(1 + y_2)^2 + 2y_2(1-y_2),$$

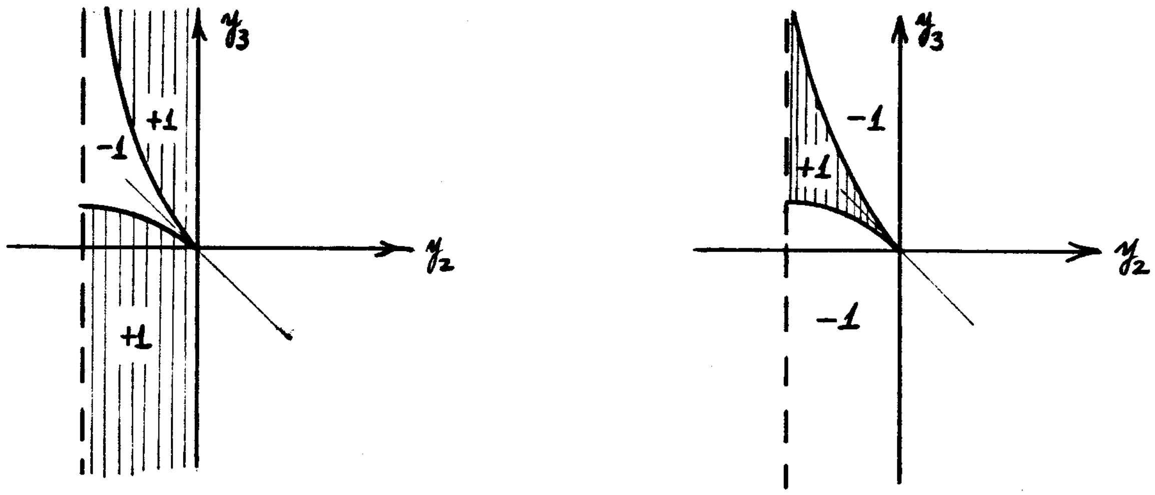
whence

$$\left.\frac{\partial f}{\partial y_3}\right|_{y_3^I} = -4(-y_2)^{3/2} < 0$$

$$\left.\frac{\partial f}{\partial y_3}\right|_{y_3^{II}} = 4(-y_2)^{3/2} > 0$$

$$f(-\tfrac{1}{2},\ 0) = \tfrac{1}{4} > 0$$

Hence sgn $f(y_2,y_3)$ and sgn$(-f)$ are as shown in Fig. 5.

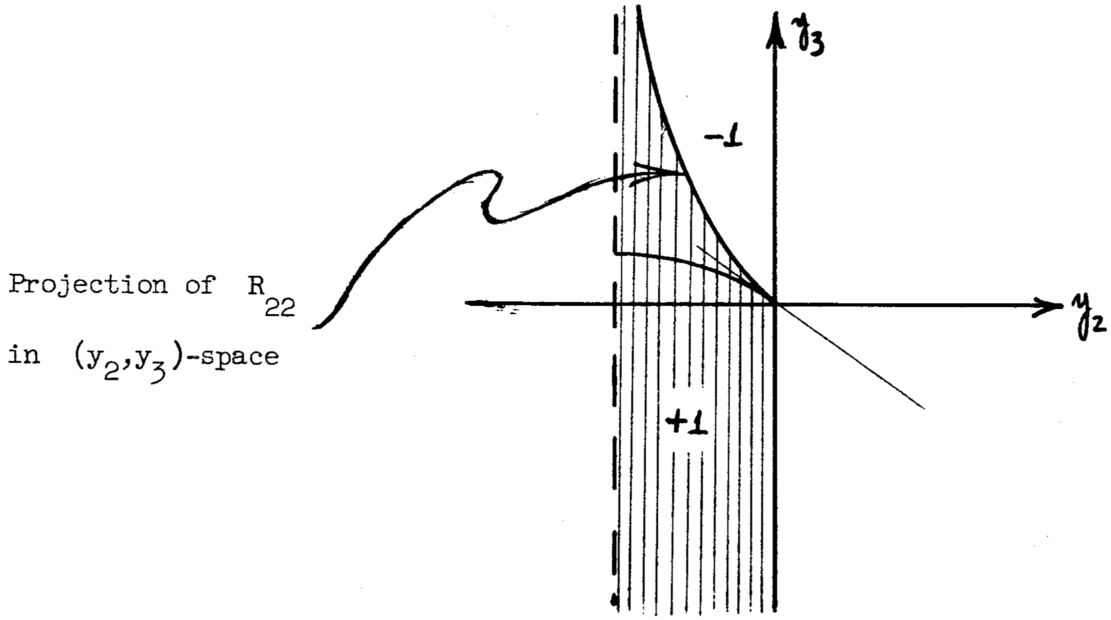


Sgn f

Sgn(-f)

FIGURE 5

Finally, $\operatorname{sgn}[-f(y_2,y_3) \cdot (y_3 - y_3^I)]$ is displayed in Fig. 6.



$\operatorname{sgn}[-f\cdot(y_3 - y_3^I)] = \operatorname{sgn}[-F(y_2,y_3)]$

FIGURE 6

Thus, the function $\operatorname{sgn}[-f\cdot(y_3 - y_3^I)]$ assigns opposite signs on the two sides of the switching surface within the set Σ. Therefore it (or its negative) could serve as a switching function within the set Σ. All that remains is to check the validity of this function for a single point within Σ. We shall assume for the moment that

$\text{sgn}[-f(y_2,y_3) \cdot (y_3 - y_3^I)]$ is indeed a correct switching function within the set Σ (more exactly, within that part of the controllable region which lies in Σ).

Let $f(y_2,y_3) \cdot (y_3 - y_3^I) = F(y_2,y_3)$. Then the switching function within Σ is simply $\text{sgn}(-F(y_2,y_3))$.

It follows by complete analogy that the switching function within Σ' is $\text{sgn}(F(z_2,z_3))$. We take sgn F rather than sgn(-F) in order to account for the fact that the value of ϵ in corresponding regions of Σ and Σ' is reversed (Fig. 7).

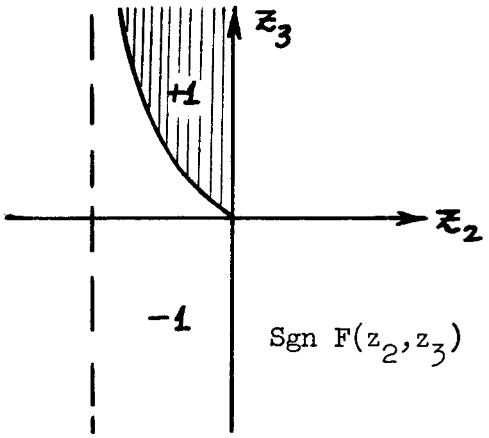


FIGURE 7

We have, by definition,

$$F(y_2,y_3) = f(y_2,y_3) \cdot (y_3 - y_3^I) \tag{25}$$

Therefore, by (24) and (20),

$$F(y_2,y_3) = \{[y_3(1 + y_2) - y_2]^2 + 4y_2y_3\} \cdot \{y_3 - \frac{y_2^2 - y_2 - 2(-y_2)^{3/2}}{(1 + y_2)^2}\} \tag{26}$$

whence

$$F(z_2,z_3) = \{[z_3(1 + z_2) - z_2]^2 + 4z_2z_3\} \cdot \{z_3 - \frac{z_2^2 - z_2 - 2(-z_2)^{3/2}}{(1 + z_2)^2}\} \tag{27}$$

Clearly

$$\begin{aligned} F(z_2,z_3) = \{[&-(-z_3)(1-(-z_2)) + (-z_2)]^2 \\ &+ 4(-z_2)(-z_3)\} \cdot \{-(-z_3) - \frac{(-z_2)^2 + (-z_2) - 2(-z_2)^{3/2}}{(1-(-z_2))^2}\} \end{aligned} \tag{28}$$

and therefore

$$F(z_2,z_3) = G(-z_2,-z_3), \tag{29}$$

where

$$G(\xi_2,\xi_3) = \{[-\xi_3(1-\xi_2) + \xi_2]^2 + 4\xi_2\xi_3\} \cdot \{-\xi_3 - \frac{\xi_2^2 + \xi_2 - 2\xi_2^{3/2}}{(1-\xi_2)^2}\} \tag{30}$$

It is obvious from the above discussion that sgn $G(-z_2,-z_3)$ is a control function within the controllable region contained in Σ'.

The reader can now easily convince himself that a control function throughout the controllable part of phase space (except for a set of three-dimensional measure zero) is given by

ϵ = sgn σ, where

$$\sigma = -(1 + \sigma_1)F(\sigma_2,\sigma_3) + (1-\sigma_1)G(\sigma_2,\sigma_3). \tag{31}$$

It seems helpful at this stage to list together the definitions of all the components which enter into the definition of σ. They are:

$$\epsilon_1 = \operatorname{sgn}(-x_1)$$

$$h_2(x_1,x_2;\eta) = -\eta + e^{-\eta\lambda x_1}(\lambda x_2+\eta)$$

$$h_3(x_1,x_2,x_3;\eta) = \eta + e^{\eta\lambda x_1}(\lambda x_3-\eta)$$

$$\epsilon_2(x_1,x_2) = -\operatorname{sgn} h_2(x_1,x_2,\epsilon_1)$$

$$\sigma_1(x_1,x_2,x_3) = \epsilon_2(x_1,x_2)$$

$$\sigma_2(x_1,x_2,x_3) = h_2(x_1,x_2;\ \sigma_1)$$

$$\sigma_3(x_1,x_2,x_3) = h_3(x_1,x_2,x_3;\ \sigma_1)$$

$$F(\xi_2,\xi_3) = \{[\xi_3(1+\xi_2)-\xi_2]^2+4\xi_2\xi_3\} \cdot \{\xi_3 - \frac{\xi_2^2-\xi_2-2(-\xi_2)^{3/2}}{(1+\xi_2)^2}\}$$

$$G(\xi_2,\xi_3) = \{[-\xi_3(1-\xi_2)+\xi_2]^2+4\xi_2\xi_3\} \cdot \{-\xi_3 - \frac{\xi_2^2+\xi_2-2\xi_2^{3/2}}{(1-\xi_2)^2}\}$$

$$\sigma = -(1+\sigma_1)F(\sigma_2,\sigma_3) + (1-\sigma_1)G(\sigma_2,\sigma_3)$$

$$\epsilon = \epsilon_3(x_1,x_2,x_3) = \operatorname{sgn} \sigma.$$

The proof above was based on the assumption that $\operatorname{sgn}[-f(y_2,y_3) \cdot (y_3 - y_3^I)]$ was the correct switching function within the set Σ, rather than $\operatorname{sgn}[+f(y_2,y_3) \cdot (y_3 - y_3^I)]$. Thus, up until now, the switching function is undetermined as to sign. This sign could be determined experimentally by simulating the system for just one set of initial values. We can also show mathematically that the sign resulting from the above assumption is indeed correct. This is done as follows:

Consider an arbitrary trajectory which meets R_{22}. Suppose its first point of intersection with R_{22} is P, where P in an interior point of R_{22}. Just after reaching P the value of ϵ is + 1. Since the value of ϵ must switch from -1 to + 1, or vice versa, at points where a trajectory first meets a switching manifold, the value of ϵ must have been -1 just before the trajectory reached P. Hence, if $\mathrm{sgn}[\pm f(y_2,y_3)\cdot(y_3-y_3^I)]$ is the true switching function in the set Σ, the $\pm$ sign must be determined so that $\pm f(y_2,y_3)$ is negative just before the trajectory reaches P. In making this assertion, we also use the fact that $y_3^{II} > y_3^I$.

From the fact that $f(y_2,y_3) = 0$ <u>at</u> P, we thus see that $\pm$ must be determined in such a way that

$$\frac{d}{dt}[\pm f(y_2,y_3)] \geq 0, \qquad (32)$$

where the differentiation is carried out along the trajectory (with

$\epsilon = -1$) through P and evaluated at P. Now the equations of motion (for $\epsilon = -1$) when expressed in terms of the y's are as follows:

$$\frac{dy_1}{dt} = -1$$

$$\frac{dy_2}{dt} = -2\lambda e^{-\lambda y_1} + 2\lambda(y_2 + 1) \qquad (33)$$

$$\frac{dy_3}{dt} = -\, 2\lambda e^{\lambda y_1} - 2\lambda(y_3 - 1)$$

The reader may verify these equations by differentiating the equations (2.6) on p. 108 of Vol. 1 of FPR and reducing the result with equations (2.4) and (2.5) on p. 107 of same.

We thus have

$$\frac{d}{dt}\,[\pm\, f(y_2,y_3)] = \pm[\frac{\partial f}{\partial y_2}\;\frac{dy_2}{dt} + \frac{\partial f}{\partial y_3}\;\frac{dy_3}{dt}\,] \qquad (34)$$

where the values of dy_2/dt and dy_3/dt are obtainable from (33). We need to carry out our argument for only one particular point P on R_{22}. A convenient choice for P is $y_1 = 0$, $y_2 = -\frac{1}{4}$, $y_3 = 1$. It is easily seen that this point does indeed lie on R_{22}, for it satisfies the two inequalities and the single equality characteristic of R_{22}. Since $f(y_2,y_3) = (y_3-y_2+ y_2y_3)^2 + 4y_2y_3$, a short calculation shows that at P

$$\frac{\partial f}{\partial y_2} = 4 \quad \text{and} \quad \frac{\partial f}{\partial y_3} = \frac{1}{2}$$

while, from (33), we find that at P

$$\frac{dy_2}{dt} = -\frac{1}{2}\lambda \quad , \quad \text{and} \quad \frac{dy_3}{dt} = -2\lambda$$

Hence, from (34), we obtain

$$\frac{d}{dt}[\pm f(y_2,y_3)]_P = \pm [4(-\tfrac{1}{2}\lambda) + \tfrac{1}{2}(-2\lambda)] = \mp 3\lambda.$$

But since $\lambda > 0$, we see at once that (32) can be satisfied only if $\mp = +$, or, in other words, only if $\pm = -$, as we wished to show.

2. Appendix to §1.

The Phase-Portrait Of System (2) In The Controllable Region Of The (y_1,y_2)-Plane.

In the region Σ^*, where $\epsilon_2 = + 1$, all trajectories are parallel to the y_1-axis. In the complementary region, where $\epsilon_2 = -1$, the trajectories form curves whose equations are given by:

$$y_2 = e^{-2\lambda y_1}(y_2(0)-1) + 2e^{-\lambda y_1} - 1 \tag{35}$$

Note that in (35) the time t has been eliminated; the argument in $y_2(0)$ is y_1, not t.

In that part of the controllable region where $\epsilon_2 = -1$, we have $0 < y_2(0) < 1$ (Fig. 3). Let $1-y_2(0) = \mu$. Then $0 < \mu < 1$ and

$$\frac{dy_2}{dy_1} = 2\lambda e^{-\lambda y_1}[\mu e^{-\lambda y_1} -1].$$

The function $g(y_1) = \mu e^{-\lambda y_1} -1$ is monotonically decreasing. It has exactly one zero at

$$y_1^* = \frac{1}{\lambda} \log \mu.$$

It is easy to see that $y_1^* < 0$, $y_1^* \to -\infty$ as $\mu \to 0$ (i.e., as $y_2(0) \to 1$) and $y_1^* \to 0$ as $\mu \to 1$ (i.e., as $y_2(0) \to 0$). Moreover,

$$\frac{d^2y_2}{dy_1^2} = 2\lambda^2 e^{-\lambda y_1}[1-2\mu e^{-\lambda y_1}],$$

so that d^2y_2/dy_1^2 is monotonically increasing and has exactly one zero at

$$y_1^{**} = \frac{1}{\lambda} \log 2\mu = y_1^* + \frac{1}{\lambda} \log 2.$$

It follows that y_1^* is a point of relative maximum while y_1^{**} is a point of inflection.

To find the point of intersection of a given trajectory with R_{11}, substitute $y_2 = 0$ in (35). One gets

$$e^{-\lambda y_1} = \frac{1 \pm \sqrt{y_2(0)}}{1 - y_2(0)} = \frac{1}{1 \mp \sqrt{y_2(0)}}$$

However, since $y_1 < 0$ at the point of intersection, we must have $e^{-\lambda y_1} > 1$ at that point. Therefore, the point $(y_1' , 0)$ at which the given trajectory intersects R_{11} is given by

$$e^{-\lambda y_1'} = \frac{1}{1 - \sqrt{y_2(0)}}$$

Clearly $y_1' \to -\infty$ as $y_2(0) \to 1$ and $y_1' \to 0$ as $y_2(0) \to 0$. The complete phase portrait is displayed in Fig. 8.

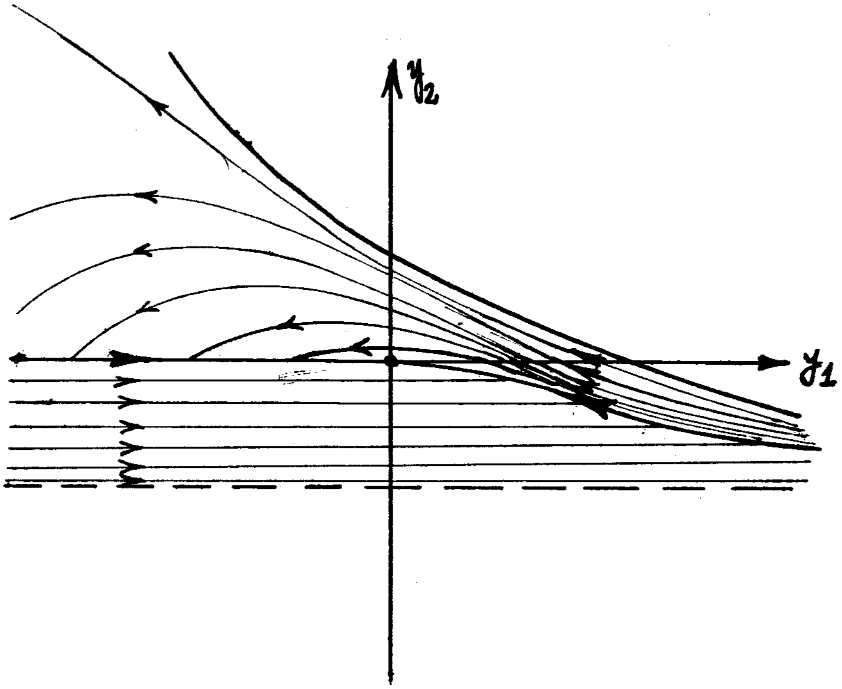


FIGURE 8

CHAPTER 16

ON A CONTROL FUNCTION FOR CONTROLLABLE LINEAR SYSTEMS WITH FOUR ZERO EIGENVALUES

1. Preliminaries

We have spent considerable effort in attacking the problem of optimal control for controllable fourth order systems having eigenvalues $0,0,\lambda, -\lambda$. The case $\lambda \neq 0$ appears to be quite difficult though not hopeless. In order to gain experience, we have considered the case $\lambda = 0$. Here our general method requires the analysis of the three-dimensional switching manifold given by (31) of FPR, Vol. 1, p. 33. In particular it is necessary to study the last equation of this chapter, which, for convenience, we reproduce here.

$$\begin{aligned} &288(72y_2^2y_4^4 - 48y_2y_3^2y_4^3 - 288y_2^4y_4^3 + 872y_2^3y_3^2y_4^2 \\ &+ 307y_2^6y_4^2 - 744y_2^2y_3^4y_4 - 425y_2^5y_3^2y_4 - 38y_2^8y_4 \\ &+ 192y_2y_3^6) + 16(2581y_2^4y_3^4 + 443y_2^7y_3^2) + 361y_2^{10} = 0. \end{aligned} \tag{1.1}$$

We are particularly interested in studying the manner in which this equation defines y_4 as a function of y_2 and y_3.

The equation is seen to be homogeneous of weight 20 in the y's if y_2 is given weight two, y_3 weight three, and y_4 weight four. Hence we may obtain a more convenient form of the equation if we set

$$\xi = \frac{12y_4}{y_2^2} \quad \text{and} \quad \zeta = \frac{4y_3^2}{y_2^3} \;, \tag{1.2}$$

where we have introduced the numerical constants 12 and 4 because they seem to make the resulting equation have smaller coefficients. If we divide (1.1) by y_2^{10} and introduce the notation defined by (1.2), we find that (1.1) is equivalent(when $y_2 \neq 0$) to

$$\xi^4-(2\zeta + 48)\xi^3 + (436\zeta + 614)\xi^2 - (1{,}116\zeta^2 + 2{,}550\zeta + 912)\xi$$
$$+(864\zeta^3 + 2{,}581\zeta^2 + 1{,}772\zeta + 361) = 0 \tag{1.3}$$

and our problem thus reduces to the study of the four roots of the fourth degree equation (1.3) in ξ as functions of ζ.

We rewrite equation (1.3) in the form

$$\xi^4 - a\xi^3 + b\xi^2 - c\xi + d = 0, \tag{1.4}$$

where

$$\begin{aligned} a &= 2\zeta + 48 \\ b &= 436\zeta + 614 \\ c &= 1{,}116\zeta^2 + 2{,}550\zeta + 912 \\ d &= 864\zeta^3 + 2{,}581\zeta^2 + 1{,}772\zeta + 361. \end{aligned} \tag{1.5}$$

We started out with the conjecture that equation (1.3), except for $\zeta = 0$, always has just two distinct real roots. A well known necessary and sufficient condition that a quartic such as (1.4) have just two distinct real roots is that its discriminant Δ be always negative. Δ may be written down in various ways. For instance,

$$\Delta = \begin{vmatrix} 1 & -a & b & -c & d & 0 & 0 \\ 0 & 1 & -a & b & -c & d & 0 \\ 0 & 0 & 1 & -a & b & -c & d \\ 4 & -3a & 2b & -c & 0 & 0 & 0 \\ 0 & 4 & -3a & 2b & -c & 0 & 0 \\ 0 & 0 & 4 & -3a & 2b & -c & 0 \\ 0 & 0 & 0 & 4 & -3a & 2b & -c \end{vmatrix}$$

$$\begin{aligned}\Delta = {} & -27a^4d^2 + 18a^3bcd - 4a^3c^3 - 4a^2b^3d + a^2b^2c^2 + 144a^2bd^2 - 6a^2c^2d \\ & -80ab^2cd + 18abc^3 - 192acd^2 + 16b^4d - 4b^3c^2 - 128b^2d^2 + 144bc^2d \\ & -27c^4 + 256d^3.\end{aligned}$$

$$\Delta = 4(\tfrac{1}{3}b^2 - ac + 4d)^3 - 27(-\tfrac{2}{27}b^3 + \tfrac{1}{3}\,abc + \tfrac{8}{3}\,bd - c^2 - a^2d)^2 .$$

Hence the question can be settled definitely in one way or another by substituting (for a,b,c,d, in any of these formulas for Δ) the expressions given by (1.5) and thus obtaining $\Delta = \Delta(\zeta)$ as a polynomial in ζ. If our conjecture were true, the equation $\Delta(\zeta) = 0$ should have had no real roots, except for a double root $\zeta = 0$, and $\Delta(\zeta)$ should have been negative for all real $\zeta \neq 0$. This, however, is not quite the case, as will be shown in the following section. The calculation of $\Delta(\zeta)$ was carried out by G. Campbell whose fortitude won our unqualified admiration, for the task turned out to be extremely laborious. It took some time to obtain $\Delta(\zeta)$ in a form free from error as indicated by a system of various checks. In the meantime we managed to run up several blind alleys. These are indicated below for the sake of completeness.

<u>First</u>, it is clear from the third formula given above for Δ that a sufficient condition that $\Delta < 0$ is that $\frac{1}{3}b^2 - ac + 4d < 0$. It is much easier to compute $\frac{1}{3}b^2 - ac + 4d$ in terms of ζ than it is to compute the full expression for Δ. It turns out to be a cubic polynomial in ζ which vanishes when $\zeta = -4.46$ approximately and which is negative when $\zeta < -4.46$. Hence equation (1.3) does have exactly two real distinct roots for $\zeta < -4.46$.

<u>Second</u>, it is obvious geometrically that (1.3) cannot have more than two real roots if the left member of (1.3) is a convex function

of ξ, i.e., if its second derivative $12\xi^2 - 6a\xi + 2b$ never changes sign. This will be the case if the discriminant of the quadratic function $6\xi^2 - 3a\xi + b$ is negative, i.e., if $9a^2 - 24b < 0$ or $3a^2 - 8b < 0$. This is found to be satisfied for values of ζ between approximately .7 and 241.9.

Third, we carried out a machine calculation of the roots of (1) for all even integral values of ζ from -500 to + 500; and it was found that in each case (except for $\zeta = 0$, of course) there were exactly 2 real roots and 2 complex roots.

None of these observations yielded sufficient information.

2. Study Of The Equation For The Three-Dimensional Switching Manifold For The System $\dot{x}_1 = \epsilon$, $\dot{x}_2 = x_1$, $\dot{x}_3 = x_2$, $\dot{x}_4 = x_3$.

The equation referred to in the section title is equation (1.1) of the present chapter. This equation was reduced by means of the substitution (1.2) to the somewhat more tractable form (1.3). Our problem thus reduces at first to the study of the four roots of the fourth degree equation (1.3) in ξ as functions of ζ.

When $\zeta = 0$, the left member of equation (1.3) may be written as a perfect square $(\xi^2 - 24\xi + 19)^2$. Hence for $\zeta = 0$, (1.3) admits two pairs of double roots $12 \pm 5\sqrt{5}$ all four roots being real.

As mentioned above, we conjectured at first that equation (1.3) would admit just two real roots and a pair of conjugate complex roots. This conjecture was supported by some elaborate numerical computations carried out on a computer and also by some theoretical work which proved that (1.3) actually does have just two real roots and a pair of conjugate complex roots as long as ζ was restricted to certain specified intervals. We now know, however, that the conjecture is false if and only if $-4 \leqq \zeta \leqq -100/27 = -3.70370370\ldots$. We also know that the curve whose equation is (1.3) has cusps at the points $\zeta = -4$, $\xi = -7$, and $\zeta = -100/27$, $\xi = -19/3$ and that the curve also has a double point at $\zeta = -125/32$, $\xi = -27/4$. On the open interval $-4 < \zeta < -125/32$, equation (1.3) has four distinct real roots; three negative roots and one positive root. The same is true for the open interval $-125/32 < \zeta < -100/27$. Notice, however, that the whole interval $-4 \leqq \zeta \leqq -100/27$ where our conjecture turns out to be false is very short. Indeed its overall length is only 8/27.

These facts were discovered and established with the help of the discriminant

$$\Delta(\zeta) = -27a^4d^2 + 18a^3bcd - 4a^3c^3 - 4a^2b^3d + a^2b^2c^2 + 144a^2bd^2$$
$$- 6a^2c^2d - 80ab^2cd + 18abc^3 - 192acd^2 + 16b^4d - 4b^3c^2$$
$$- 128b^2d^2 + 144bc^2d - 27c^4 + 256d^3, \qquad (2.1)$$

which has already been introduced above. Here $a,b,c,d,$ are the same as in (1.5)

A necessary and sufficient condition that (1.4) have just two real distinct roots is that $\Delta < 0$ and it is also well known that Δ vanishes if and only if (1.4) has a multiple root. Hence we calculated Δ as a polynomial in ζ by substituting in (2.1) the values of a,b,c,d given by (1.5). Δ must, of course, vanish at $\zeta = 0$ to the second order because, as we have already pointed out, for $\zeta = 0$ our equation (1.3) has two double roots. Hence we are justified in writing Δ in the form

$$\Delta = -\zeta^2F(\zeta) \qquad (2.2)$$

It required a very stupendous calculation to find the polynomial $F(\zeta)$. But we eventually found that

$$F(\zeta) = 322{,}486{,}272\zeta^8 + 9{,}972{,}440{,}064\zeta^7 + 134{,}895{,}988{,}656\zeta^6$$

$$+ 1{,}042{,}527{,}831{,}808\zeta^5 + 5{,}034{,}833{,}427{,}200\zeta^4$$

$$+ 15{,}559{,}336{,}960{,}000\zeta^3 + 30{,}047{,}296{,}000{,}000\zeta^2$$

$$+ 33{,}152{,}000{,}000{,}000\zeta + 16{,}000{,}000{,}000{,}000 \qquad (2.3)$$

After still more harrowing adventures we found that $F(\zeta)$ could be factored into linear factors. Namely,

$$F(\zeta) = 16t^3(32t - 3)^2(27t-8)^3 \qquad (2.4)$$

where

$$t = \zeta + 4 \qquad (2.5)$$

That is,

$$F(\zeta) = 16(\zeta + 4)^3(32\zeta + 125)^2(27\zeta + 100)^3 \qquad (2.6)$$

It seems undesirable to burden the reader with the many details leading to the discovery of the factorization of F as exhibited in (2.4) or (2.6). It is indeed burdensome for him to verify the correctness of (2.6) a posteriori by multiplying out the factors but not as much so as a detailed discussion would be as to how the factors were discovered in the first place.

It is seen immediately from (2.6) that (1.3) has multiple roots when $\zeta = -4$, $\zeta = -\frac{125}{32}$, $\zeta = -\frac{100}{27}$. These multiple roots all turn out to be double roots and they are, respectively, $\xi = -7$, $\xi = -\frac{27}{4}$, and $\xi = -\frac{19}{3}$.

In order to verify these facts it is desirable to use (2.5) and the further substitution

$$\lambda = 7 + \xi - \frac{5}{2}t \tag{2.7}$$

to write (1.3) in the somewhat simpler form

$$\lambda^4 + (8t-68)\lambda^3 + (\frac{45}{2}t^2 - 32t + 4)\lambda^2 + (25t^3 - t^2)\lambda + (\frac{125}{16}t^4 - t^3) = 0 \tag{2.8}$$

Incidentally, by neglecting all terms in (2.8), except those of lowest order, we obtain $\lambda = \pm \frac{1}{2}\sqrt{t^3} + \ldots$, which shows that the curve whose equation is given by (2.8) has a cusp at the origin. This corresponds to a cusp at $\zeta = -4$, $\xi = -7$ in the curve whose equation is (1.3). The latter curve also has a cusp at $\zeta = -\frac{100}{27}$, $\xi = -\frac{19}{3}$, as it would be possible to establish in a similar manner. But this fact can also be deduced more readily by inspection of (2.6).

Details are left to the reader.

In order to study the curve given by (1.3) for large values of ξ and ζ it may be noticed that the left hand side of (1.3) is almost divisible by $\xi - 2\zeta - 1$. In fact, if we attempt to carry out such a division we get a quotient

$$Q(\xi,\zeta) = \xi^3 - 47\xi^2 + (342\zeta + 567)\xi - (432\zeta^2 + 1{,}074\zeta + 345) \tag{2.9}$$

and a remainder

$$R(\zeta) = (\zeta + 4)^2 \tag{2.10}$$

Hence, if the left member of (1.3) is denoted by $F(\xi,\zeta)$, we have

$$F(\xi,\zeta) = Q(\xi,\zeta)(\xi - 2\zeta - 1) + (\zeta + 4)^2 \tag{2.11}$$

Thus at any point on the straight line $\xi = 2\zeta + 1$, or on the curve whose equation is $Q(\xi,\zeta) = 0$, we must have $F(\xi,\zeta) = (\zeta + 4)^2 \geqq 0$, with the equality sign holding if and only if $\zeta = -4$. It follows that for those values of ζ where the equation (1.3) has just two real distinct roots (which it does except for $-4 \leqq \zeta \leqq -100/27$ and for $\zeta = 0$) the points of the curve $F(\xi,\zeta) = 0$ must all lie either completely below the straight line $\xi = 2\zeta + 1$, as in the case for $\zeta > 0$, or completely above the straight line $\xi = 2\zeta + 1$, as in the case for $\zeta < -4$. A similar statement may be made with regard to the curve $Q(\xi,\zeta) = 0$. But it is best to confine attention not only to those values of ζ for which $F(\xi,\zeta) = 0$ has just two real roots but also for those values of ζ for which the cubic $Q(\xi,\zeta) = 0$ has just one real root. It is easy to plot the curve $Q(\xi,\zeta) = 0$ because, although it is cubic in ξ, it is only quadratic in ζ. Thus we may use the quadratic formula to solve $Q(\xi,\zeta) = 0$ for ζ in terms of ξ. The result is

$$\zeta = \frac{1}{144}\left[57\xi - 179 \pm \sqrt{p(\xi)}\right] \tag{2.12}$$

where $p(\xi) = 48\xi^3 + 993\xi^2 + 6810\xi + 15481$. Since $p(\xi) = 0$ has three roots at approximately $\xi = -7.8$, -6.9, and -6.0, we see that the curve $Q(\xi,\zeta) = 0$ consists of a "main" branch (reaching from a point where $\xi = -6.0$, approximately, to points where $\xi \to +\infty$) and of an isolated tiny oval (extending from a point where $\xi = -7.8$ to a point where $\xi = -6.9$ approximately). If we restrict attention to values of $\zeta < -5$ or $> +1$, we not only eliminate all necessity for considering this little oval, but we also eliminate the points near $\zeta = 0$ where the equation $Q(\xi,\zeta) = 0$ has three instead of only one real root. See figure 2.3

For $\zeta > 1$, then, the curve $F(\xi,\zeta) = 0$ lies completely above the curve $Q(\xi,\zeta) = 0$. For $\zeta < -5$, the curve $F(\xi,\zeta) = 0$ lies completely below the curve $Q(\xi,\zeta) = 0$.

In order to get the somewhat more exact information needed in the next section we must refine the above argument to a certain extent. What we now do is to divide $F(\xi,\zeta)$ by $(\xi - 2\zeta - 1 - \sigma)$ thus obtaining a quotient

$$Q(\xi,\zeta;\sigma) = \xi^3 + (\sigma-47)\xi^2 + [(\sigma^2-46\sigma + 567) + (2\sigma + 342)\zeta]\xi$$
$$+ [(\sigma^3 - 45\sigma^2 + 521\sigma-345) + (4\sigma^2 + 252\sigma-1074)\zeta + (4\sigma - 432)\zeta^2] \tag{2.13}$$

and a remainder

$$R(\zeta,\sigma) \equiv (\zeta + 4)^2 + \sigma[8\zeta^3 + (12\sigma + 76)\zeta^2 + (6\sigma^2 + 166\sigma + 220)\zeta$$
$$+ (\sigma^3 - 44\sigma^2 + 476\sigma + 176)] \tag{2.14}$$

so that

$$F(\xi,\zeta) = Q(\xi,\zeta;\sigma)(\xi - 2\zeta - 1 - \sigma) + R(\zeta,\sigma) \tag{2.15}$$

It is to be observed that (2.11) is a special case of (2.15) with $\sigma = 0$.

It is also important to observe that although $R(\zeta,0)$ is always positive, it is also possible to make $R(\zeta; \sigma)$ negative for sufficiently large $|\zeta|$ simply by choosing the sign of σ to be opposite to the sign of ζ. This is because the term of highest order in the expression for $R(\zeta; \sigma)$ is $8\sigma\zeta^3$. Thus, for every $\epsilon > 0$, it is possible to choose σ such that $|\sigma| < \epsilon$ and such that $R(\zeta; \sigma)$ is negative for ζ sufficiently large in absolute value and with the proper sign. If for such a ζ we choose a point (ξ_1,ζ) which lies on either the straight line $\xi = 2\zeta + 1 + \sigma$ or on the curve $Q(\xi,\zeta;\sigma) = 0$, we must have $F(\xi_1,\zeta) = R(\zeta;\sigma) < 0$, whereas, of course, if the point (ξ_0,ζ) lies on either the straight line $\xi = 2\zeta + 1$ or on the curve $Q(\xi,\zeta;\sigma) = 0$, we must have $F(\xi_0,\zeta) = R(\zeta;0) > 0$. Hence there must be a number ξ between ξ_0 and ξ_1 such that $F(\xi,\zeta) = 0$, that is, such that the point (ξ,ζ) lies on the curve $F(\xi,\zeta) = 0$. This amounts to saying that for ζ sufficiently large in absolute value and with proper sign there is a branch of the curve $F(\xi,\zeta) = 0$ which lies between the two straight lines $\xi = 2\zeta + 1$ and $\xi = 2\zeta + 1 + \sigma$ and also a branch of the curve $F(\xi,\zeta) = 0$ which lies between the two cubics $Q(\xi,\zeta) = 0$ and $Q(\xi,\zeta;\sigma) = 0$.

In the former case, as $|\zeta| \to \infty$, we have $\xi_1 - 2\zeta = 1 + \sigma$ and $\xi_0 - 2\zeta = 1$. Since $|\sigma|$ may be taken arbitrarily small, this means that $\xi - 2\zeta \to 1$ as the point (ξ,ζ) recedes indefinitely along the branch of the curve $F(\xi,\zeta) = 0$, which lies between these two straight lines.

In the latter case, when (ξ_1,ζ) and (ξ_0,ζ) are points of $Q(\xi,\zeta;\sigma) = 0$ and $Q(\xi,\zeta) = 0$, respectively, we see from (1.14) that

$$\lim_{\zeta \to \pm\infty} \frac{\zeta}{\xi_0^{3/2}} = \pm \frac{\sqrt{3}}{36}$$

We also have an equation like (2.12) which applies to $Q(\xi,\zeta,\sigma) = 0$ instead of to $Q(\xi,\zeta;0) = 0$ and which can be written down when we solve the equation $Q(\xi,\zeta;\sigma) = 0$ for ζ in terms of ξ by use of the quadratic formula. We are thus enabled to prove that

$$\lim_{\zeta \to \pm\infty} \frac{\zeta}{\xi_1^{3/2}} = \pm \frac{1}{2\sqrt{108-\sigma}}$$

Taking ζ to be positive and σ to be negative and remembering that a branch of the curve $F(\xi,\zeta) = 0$ lies above the curve $Q(\xi,\zeta;\ 0) = 0$

and (for sufficiently large ζ) below the curve $Q(\xi,\zeta;\sigma) = 0$, so that

$$\frac{\zeta}{\xi_0^{3/2}} > \frac{\zeta}{\xi^{3/2}} > \frac{\zeta}{\xi_1^{3/2}} ,$$

we see that, on this branch of the curve $F(\xi,\zeta) = 0$,

$$\limsup_{\zeta\to\infty} \frac{\zeta}{\xi^{3/2}} \leq \lim_{\zeta\to\infty} \frac{\zeta}{\xi_0^{3/2}} = \frac{\sqrt{3}}{36} \tag{2.16}$$

and that

$$\liminf_{\zeta\to\infty} \frac{\zeta}{\xi^{3/2}} \geq \lim_{\zeta\to\infty} \frac{\zeta}{\xi_1^{3/2}} = \frac{1}{2\sqrt{108-\sigma}}$$

Since this last relation holds for all negative σ, we find, by letting σ approach zero, that

$$\liminf_{\zeta\to\infty} \frac{\zeta}{\xi^{3/2}} \geq \lim_{\sigma\to 0} \frac{1}{2\sqrt{108-\sigma}} = \frac{\sqrt{3}}{36} \tag{2.17}$$

Hence, from (2.16) and (2.17), we see that $\lim\limits_{\zeta \to \infty} \frac{\zeta}{\xi^{3/2}}$ exists and is equal to $\frac{\sqrt{3}}{36}$.

Similarly, by taking ζ to be negative and σ positive, we can show that on another branch of the curve $F(\xi,\zeta) = 0$, we have

$$\lim_{\zeta \to -\infty} \frac{\zeta}{\xi^{3/2}} = -\frac{\sqrt{3}}{36}$$

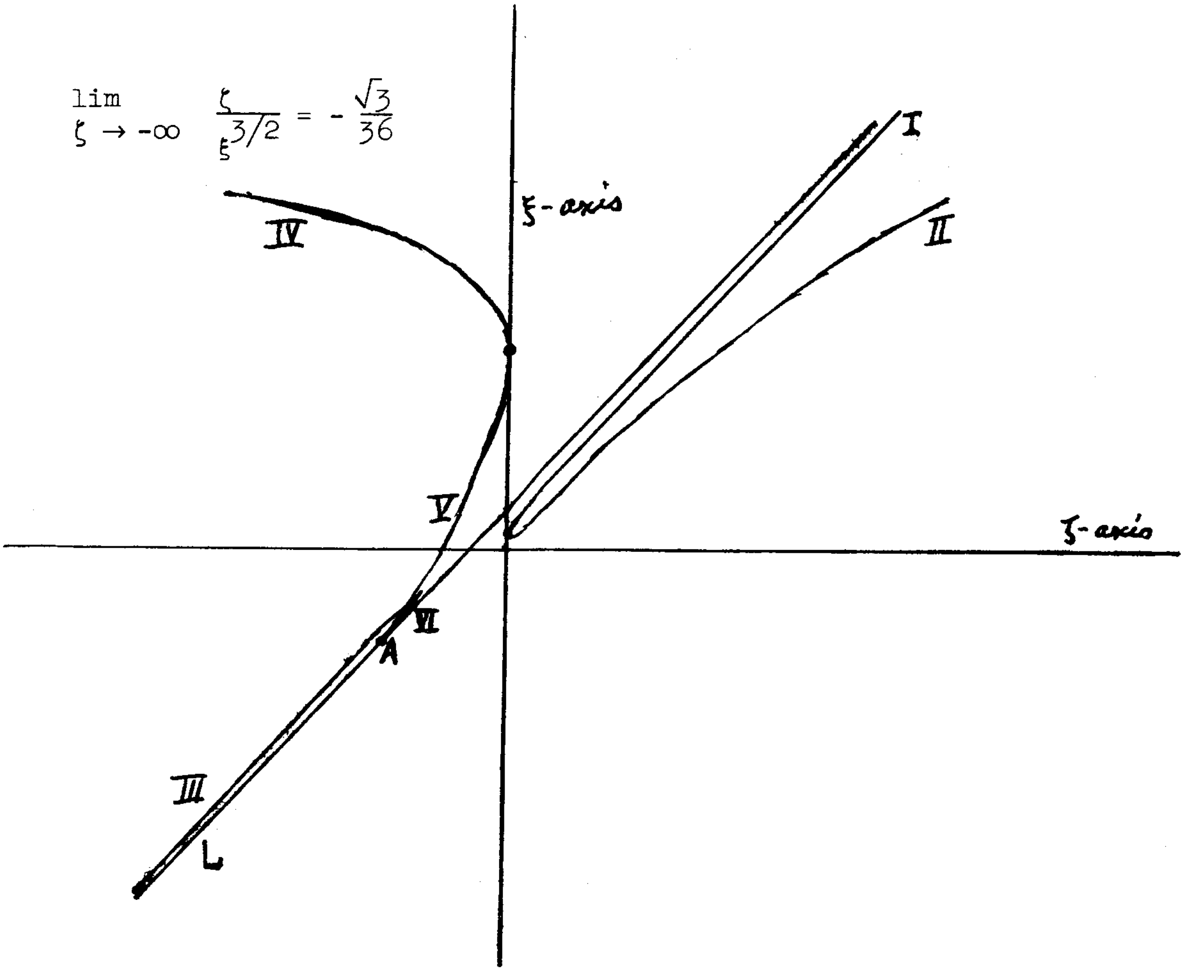


FIGURE 2.1

This figure (2.1) illustrates the main features of the curve whose equation is (1.3). The scale on the two axes is, however, slightly distorted so as to make the picture reasonably artistic. Actually the straight line, which is an asymptote to two branches of the curve, should have slope 2 instead of 1. The detailed behavior of the curve near the point A, whose coordinates are ($\xi = -7$, $\zeta = -4$), can not be depicted on such a small scale. It is shown in Figure 2.2

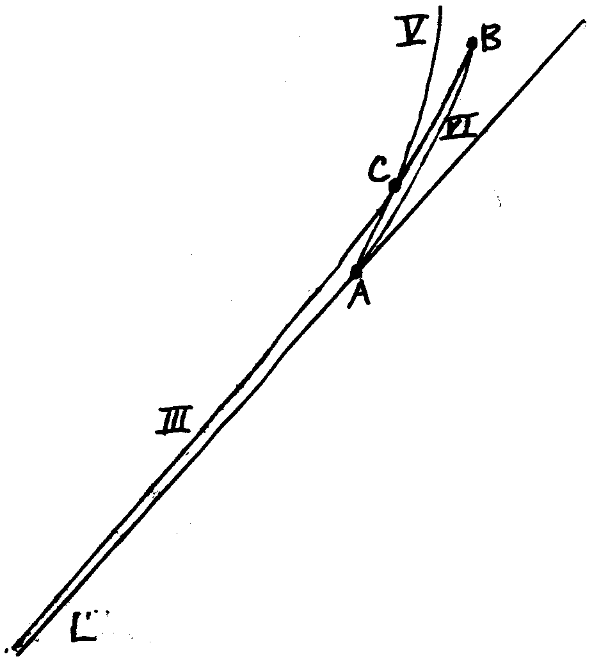


FIGURE 2.2

This figure illustrates the cusps at the points A and B and the double point at C. The straight line segment marked L represents a small piece of the straight line in Figure 2.1, also marked L . The three points A,B,C are extremely close to each other on the scale of Figure 2.1.

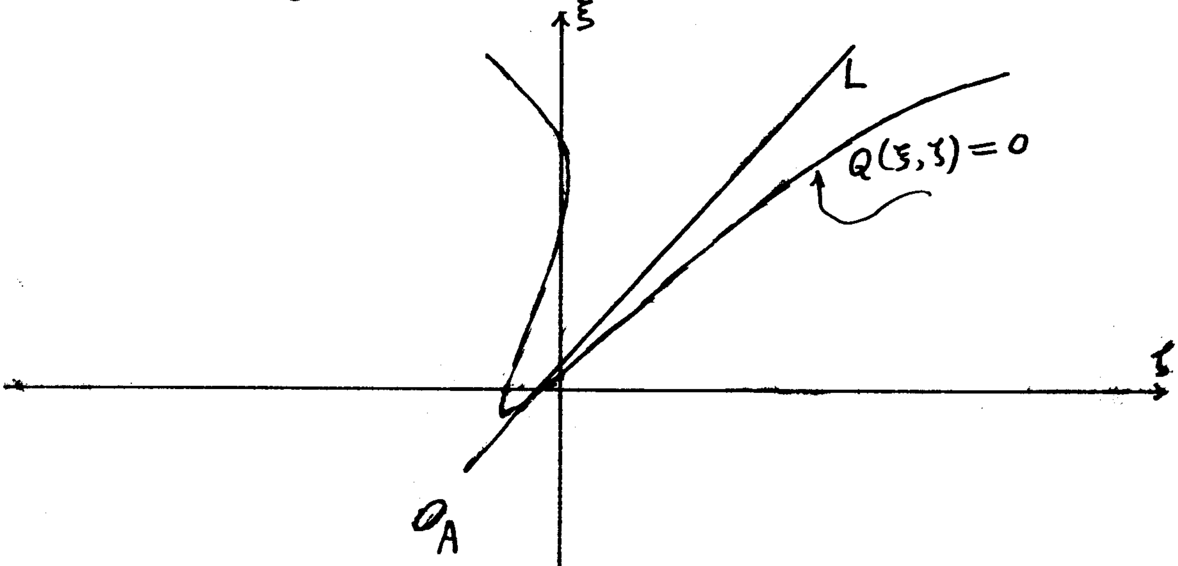


FIGURE 2.3

This figure illustrates the curve $Q(\xi,\zeta) = 0$, and the straight line $\xi = 2\zeta + 1$ marked L, to which the curve $F(\xi,\eta) = 0$, shown in Figures 2.1 and 2.2 is asymptotic. Notice the little oval near the point A. No attempt is made to maintain a thoroughly consistent scale.

3. Identification Of The Proper Leaf R_{31}

Our object is to obtain a closed form control law for the system

$$S_4: \dot{x}_1 = \epsilon, \ \dot{x}_2 = x_1, \ \dot{x}_3 = x_2, \ \dot{x}_4 = x_3, \quad \epsilon = \pm 1.$$

In Chapter 14 we outlined a general scheme for the solution of this type of problem. The procedure, as applied to the present problem, may be summarized as follows:

Consider the system

$$S_3: \ \dot{x}_1 = \epsilon, \ \ \dot{x}_2 = x_1, \ \ \dot{x}_3 = x_2, \quad \epsilon = \pm 1$$

and assume that a closed form control law $\epsilon = \epsilon_3(x)$ for the system S_3 is known (it is — see Chapter 13). Associated with the system S_4 are two sets of auxiliary variables (y_1, y_2, y_3, y_4) and (z_1, z_2, z_3, z_4) defined in accordance with our general theory (Chapters 1 and 4 in Vol. 1). We assume that (y_1, y_2, y_3, y_4) and (z_1, z_2, z_3, z_4) are so chosen that (y_1, y_2, y_3) and (z_1, z_2, z_3) form an appropriate set of auxiliary variables for the system S_3.

The transformation from the space of (x_1,x_2,x_3) to the space of (y_1,y_2,y_3) is well defined — and, in fact, invertible — in a neighborhood of the origin. The function $\epsilon_3(x)$ may therefore be regarded as a function of (y_1,y_2,y_3) in a neighborhood of the origin of the space of (y_1,y_2,y_3). Let Σ^* be the set of all those points in the space of (y_1,y_2,y_3) at which the function ϵ_3 takes on the value of $+1$. Let Σ be the subset of the space of (y_1,y_2,y_3,y_4) obtained by taking the Cartesian product of Σ^* with the y_4-axis. Denoting the y_4-axis by Y_4, the set Σ may be represented symbolically by $\Sigma^* \times Y_4$.

The switching surface of system S_4 is a three-dimensional manifold denoted by R_3. It is composed of two leaves, R_{31} and R_{32}, in accordance with our general theory (see Chapter 1, Vol. 1). We have shown (see Chapter 14) that R_{31}, when imbedded in the space of (y_1,y_2,y_3,y_4) be means of the transformation which carries the x's into the y's , is wholly contained within the set Σ. The leaf R_{31}, when so imbedded, has the following properties: (1) it is a cylindrical surface, parallel to the y_1-axis; (2) it separates Σ into two distinct parts; (3) its boundary lies on the boundary of Σ. In the sequel, R_{31} will always be conceived of as imbedded in the set Σ.

The crux of our method consists in the fact that a closed control law for the system S_4 can always be derived from a control law valid solely in the set Σ. It is therefore sufficient to restrict one's attention to the set Σ and ignore the remainder of phase space. Within Σ the problem reduces somewhat further. Finding a control function in Σ is equivalent to finding a function $F(y_1,y_2,y_3,y_4)$, defined throughout Σ, which is positive on one side of R_{31} and negative on the other. We attempt to construct a function F by making use of the equation and inequalities defining R_{31}. However, since R_{31} is parallel to the y_1-axis, the function $F(y_1,y_2,y_3,y_4)$ must necessarily be independent of y_1. In other words, it is sufficient to restrict one's attention to the projections of Σ and R_{31} into the space of (y_2,y_3,y_4). We denote these projections by Σ^P and R_{31}^P , respectively. Our problem, then, is to construct a function $F(y_2,y_3,y_4)$, defined throughout Σ^P, which is positive on one side of R_{31}^P and negative on the other.

But first we must make an exact identification of the proper leaf R_{31}, or, which is the same, the proper projection R_{31}^P . This identification will be pursued throughout the remainder of the present section.

The leaf R_{31} is characterized by one equation and three inequalities. These have been computed before (FPR, Vol. 1, pp.33-34). We repeat them here for the sake of convenience:

$$R_{31}: \begin{cases} E < 0 \\ -\frac{B}{A} > E \\ y_1 < -\frac{B}{A} \\ AD\text{-}BC = 0, \end{cases} \qquad (3.1)$$

where

$$A = 12y_3[-15y_2^3 + 140y_2y_4 - 96y_3^2]$$

$$B = 836y_2^2y_3^2 + 95y_2^5 - 1440y_4y_2^3 + 720y_2y_4^2 - 576y_4y_3^2$$

$$E = \frac{y_3 - 2y_2(\frac{B}{A}) - (\frac{B}{A})^3}{y_2 + (\frac{B}{A})^2}$$

$$\begin{aligned} AD - BC = {} & 288(72y_2^2y_4^4 - 48y_2y_3^2y_4^3 - 288y_2^4y_4^3 + 872y_2^3y_3^2y_4^2 \\ & + 307y_2^6y_4^2 - 744y_2^2y_3^4y_4 - 425y_2^5y_3^2y_4 - 38y_2^8y_4 \\ & + 192y_2y_3^6) + 16(2581y_2^4y_3^4 + 443y_2^7y_3^2) + 361y_2^{10} \end{aligned}$$

The inequality $E < -\frac{B}{A}$ may be simplified as follows:

One has

$$E + \frac{B}{A} < 0,$$

whence

$$\frac{y_3 - 2y_2(\frac{B}{A}) - (\frac{B}{A})^3}{y_2 + (\frac{B}{A})^2} + (\frac{B}{A}) < 0$$

Therefore

$$\frac{y_3 - 2y_2(\frac{B}{A}) - (\frac{B}{A})^3 + y_2(\frac{B}{A}) + (\frac{B}{A})^3}{y_2 + (\frac{B}{A})^2} < 0$$

or

$$\frac{y_3 - y_2(\frac{B}{A})}{y_2 + (\frac{B}{A})^2} < 0 \qquad (3.2)$$

The inequality $E < 0$ may also be simplified. One has

$$\frac{y_3 - 2y_2(\frac{B}{A}) - (\frac{B}{A})^3}{y_2 + (\frac{B}{A})^2} < 0$$

whence

$$\frac{y_3 - y_2(\frac{B}{A})}{y_2 + (\frac{B}{A})^2} - \frac{y_2(\frac{B}{A}) + (\frac{B}{A})^3}{y_2 + (\frac{B}{A})^2} = \frac{y_3 - y_2(\frac{B}{A})}{y_2 + (\frac{B}{A})^2} - \frac{B}{A} < 0 \qquad (3.3)$$

Using (3.2) and (3.3) we may therefore replace (3.1) by

$$R_{31}: \begin{cases} \varphi < 0 \\ \varphi - \frac{B}{A} < 0 \\ y_1 < -\frac{B}{A} \\ AD-BC = 0 \end{cases} \qquad (3.4)$$

where

$$\varphi = \frac{y_3 - y_2(\frac{B}{A})}{y_2 + (\frac{B}{A})^2} \qquad (3.5)$$

The functions A,B,C,D and φ are all independent of y_1. Hence the projection R_{31}^{P} is characterized by the equation

$$AD-BC = 0 \qquad (3.6)$$

and the inequalities

$$\varphi < 0 \tag{3.7}$$

$$\varphi - \frac{B}{A} < 0 \tag{3.8}$$

where φ is as defined in (3.5).

Let ξ and ζ be the variables defined above, namely

$$\xi = \frac{12y_4}{y_2^2} \; ; \qquad \zeta = \frac{4y_3^2}{y_2^3}$$

and let

$$\Phi = \frac{y_2 B}{y_3 A}$$

Then (3.7) becomes

$$\frac{y_3(1-\Phi)}{y_2(1 + \frac{\zeta}{4}\Phi^2)} < 0 \, , \tag{3.9}$$

whereas (3.8) becomes

$$\frac{y_3}{y_2}\,\frac{1-\Phi}{1 + \frac{\zeta}{4}\Phi^2} - \Phi < 0 \tag{3.10}$$

where

$$\Phi = \frac{y_2 B}{y_3 A} = \frac{209\zeta + 95 - 120\xi + 5\xi^2 - 12\xi\zeta}{3\zeta\{-15 + \frac{35}{3} - 24\zeta\}} \tag{3.11}$$

We shall use (3.9), (3.10) and (3.11) to distinguish between the proper and spurious leaves defined by (3.6).

As we already know from previous sections, equation (3.6), when expressed in terms of ζ,ξ gives rise to six real branches in the (ζ,ξ) plane. We shall refer to them as branches I through VI in accordance with Figures 2.1 and 2.2.

Lemma 1. $\Phi(\zeta,\xi) \to 2$ as $\zeta \to +\infty$ on branch I.

Proof: We know from the discussion in previous paragraphs that $[\xi-(2\zeta + 1)] \to 0$ as $\zeta \to +\infty$ on branch I. Hence, on this branch, we have from (3.11):

$$\lim_{\zeta\to+\infty} \Phi(\zeta,\xi) = \lim_{\zeta\to+\infty} \frac{209\zeta + 95 - 120(2\zeta + 1) + 5(2\zeta + 1)^2 - 12\zeta(2\zeta + 1)}{\zeta(-45 + 35(2\zeta + 1) - 72\zeta)}$$

$$= \lim_{\zeta\to+\infty} \frac{-4\zeta^2 + P_1(\zeta)}{-2\zeta^2 + Q_1(\zeta)}$$

where $P_1(\zeta)$ and $Q_1(\zeta)$ are polynomials of degree 1 in ζ. It clearly follows that

$$\lim_{\zeta\to+\infty} \Phi(\zeta,\xi) = 2$$

along branch I.

Lemma 2. $\Phi(\zeta,\xi) \to 2$ as $\zeta \to -\infty$ on branch III.

Proof: The proof is the same.

Lemma 3. $\Phi(\zeta,\xi) \to 0$ as $\zeta \to +\infty$ [$\zeta \to -\infty$] on branch II [branch IV].

Proof: It has already been shown that on branch II the value of ζ tends asymptotically to $\frac{\sqrt{3}}{36}\ \xi^{3/2}$ as ζ and ξ tend to $+\infty$. In other words, on branch II,

$$\lim_{\zeta \to +\infty} \frac{\zeta}{\xi^{3/2}} = \frac{\sqrt{3}}{36}$$

We may therefore write

$$\zeta = (1 + \omega)\frac{\sqrt{3}}{36}\ \xi^{3/2} \tag{3.12}$$

where $\omega \to 0$ as $\zeta \to +\infty$ on branch II. Substitution of (3.12) into (3.11) yields a rational function of $\xi^{1/2}$ whose numerator is of degree five in $\xi^{1/2}$ whereas its denominator is of degree six in $\xi^{1/2}$. Since $\xi \to +\infty$ $\zeta \to +\infty$ on branch II, the stated result follows.

The proof for branch IV is analogous.

A. The case when $\zeta > 0$.

When $\zeta > 0$ one has $4y_3^2/y_2^3 > 0$, whence $y_2 > 0$. Hence the half plane $\zeta > 0$ corresponds to the half space $y_2 > 0$ in the space of (y_2, y_3, y_4). We shall consider this half-space as divided into two quadrants, namely (i) $y_2 > 0$, $y_3 > 0$; (ii) $y_2 > 0$, $y_3 < 0$.

(i) $y_2 > 0, \ y_3 > 0$

We have $\zeta > 0$, whence $1 + \frac{\zeta}{4} \Phi^2 > 0$. Moreover, $y_3/y_2 > 0$. Hence (3.9) is equivalent to $1-\Phi < 0$, or simply

$$\Phi > 1 \tag{3.13}$$

We note, furthermore, that if $y_2 > 0$, $y_3 > 0$ and (3.13) is satisfied then (3.10) is automatically satisfied. It follows that in the quadrant $y_2 > 0$, $y_3 > 0$ the two inequalities (3.9), (3.10) may be replaced by the single inequality (3.13).

Lemma 4. In the quadrant $y_2 > 0$, $y_3 > 0$, points corresponding to branch I with a sufficiently large ζ, satisfy (3.13).

Proof. Lemma 1.

Lemma 5. In the quadrant $y_2 > 0$, $y_3 > 0$, points corresponding to branch II with a sufficiently large ζ, violate (3.13).

Proof: Lemma 3.

Theorem 1. The subset of R_{31}^{P} which lies in the quadrant $y_2 > 0$, $y_3 > 0$ is the locus of all points (in this quadrant) which correspond to branch I. The locus of all points, which lie in the quadrant $y_2 > 0$, $y_3 > 0$ and which correspond to branch II, is spurious.

Proof: Every point of R_{31}^{P} must satisfy equation (3.6) and will, therefore, correspond to a point on one of the branches depicted in Figure 2.1. Since $y_2 > 0$ in the quadrant under consideration, ζ is positive whence every point of R_{31}^{P} for which $y_2 > 0$ and $y_3 > 0$ must either correspond to a point on branch I or to a point on branch II. Every such point of R_{31}^{P} will also satisfy the inequalities (3.7) and (3.8) or, equivalently, (3.10) and (3.11). The last two inequalities have been shown to be equivalent, in the quadrant $y_2 > 0$, $y_3 > 0$, to the single inequality (3.13). It follows that every point of R_{31}^{P} in the quadrant $y_2 > 0$, $y_3 > 0$ corresponds to a point on branch I, or branch II, which satisfies (3.13).

Conversely, every point in the quadrant $y_2 > 0$, $y_3 > 0$ which corresponds to a point on either branch I or branch II and which satisfies (3.13), lies on R^P_{31}. It follows that the subset of R^P_{31} which lies in the quadrant $y_2 > 0$, $y_3 > 0$ is the locus of all points (in this quadrant) which correspond to points on the branches I, II and which satisfy (3.13).

It is a consequence of the above remarks, as well as Lemma 1, that the locus of all points corresponding to branch I with sufficiently large ζ is contained in R^P_{31}. Similarly, it follows from Lemma 3 that the locus of all points corresponding to branch II with sufficiently large ζ is spurious.

Let Z be the set of all $\zeta > 0$ such that the points (ζ, ξ_I) lying in branch I does not correspond to points (in the quadrant $y_2 > 0$, $y_3 > 0$) which are contained in R^P_{31}. Let ζ_0 be the least upper bound of Z. We know from the preceding paragraph that ζ_0 is finite. Let ξ_0 be the value of ξ, on branch I, corresponding to ζ_0.

The pair (ζ_0, ξ_0) gives rise, in the (y_2, y_3, y_4)-space to two curves. They are formed by the intersection of the two surfaces

$$4y_3^2 - \zeta_0 y_2^3 = 0, \quad \zeta_0 > 0$$

and

$$12y_4 - \xi_o y_2^2 = 0, \quad \xi_o > 0.$$

One of these curves lies in the quadrant $y_2 > 0$, $y_3 < 0$. The other lies in the quadrant under consideration, namely $y_2 > 0$, $y_3 > 0$.

It is clear from the definition of (ζ_o, ξ_o) that the curve

$$\Gamma : 4y_3^2 - \zeta_o y_2^3 = 0, \quad 12y_4 - \xi_o y_2^2 = 0, \quad y_3 > 0 \tag{3.14}$$

forms an edge of R_{31}^P. This edge may or may not be contained in R_{31}^P depending on whether $\zeta_o \in Z$ or $\zeta_o \notin Z$, respectively. Recalling the definition of R_{31}^P we must conclude that there exists a cylindrical sheet C, parallel to the y_1-axis, whose projection in the (y_2, y_3, y_4)-space is Γ, and which forms an edge of R_{31}.

One has:

$(y_2^o, y_3^o, y_4^o) \in \Sigma^P <=>$

$<=>$ there exists a y_1^* such that $(y_1^*, y_2^o, y_3^o, y_4^o) \in \Sigma$

$<=>$ there exists a y_1^* such that $(y_1^*, y_2^o, y_3^o) \in \Sigma^*$

$<=>$ $(y_2^o, y_3^o) \in M$,

where M is the unshaded region depicted in Figure 3.1. The last equivalence follows from the definition of Σ^*, as well as the detailed investigation of the switching surface of S_3 as rendered in Chapter 11 of this Final Progress Report. (The reader who wishes to refer to that chapter should note that here we have replaced the z's by y's throughout). It follows that the quadrant $y_2 > 0$, $y_3 > 0$ in the space of (y_2, y_3, y_4) lies wholly within the interior of Σ^P.

The value of ζ_o was shown to be finite. Hence Γ cannot degenerate into a curve which lies in the plane $y_2 = 0$. In other words, the curve Γ except for the point $y_2 = y_3 = y_4 = 0$ lies properly within the interior of Σ^P.

If $\zeta_o > 0$ then Γ lies within the interior of the quadrant $y_2 > 0$, $y_3 > 0$, whence C is an edge of R_{31} lying within the interior of Σ.

In Chapter 12 of this Final Progress Report it was shown that R_3 is homeomorphic to a three-dimensional disk. R_{31} and R_{32} are joined along a common edge which lies on the boundary of Σ. In particular, R_{31} cannot have an edge which lies in the interior of Σ. It follows that $\zeta_o > 0$. Hence every point lying on branch I must correspond (in the quadrant $y_2 > 0$, $y_3 > 0$) only to points which are properly contained in R_{31}^P.

We have already shown that the locus of all points in the space of (y_2,y_3,y_4), $y_2 > 0$, $y_3 > 0$, which correspond to points on branch II with sufficiently large ζ is a spurious locus. It is now claimed that the statement holds true throughout branch II, without restriction to large values of ζ. The proof is analogous to the above.

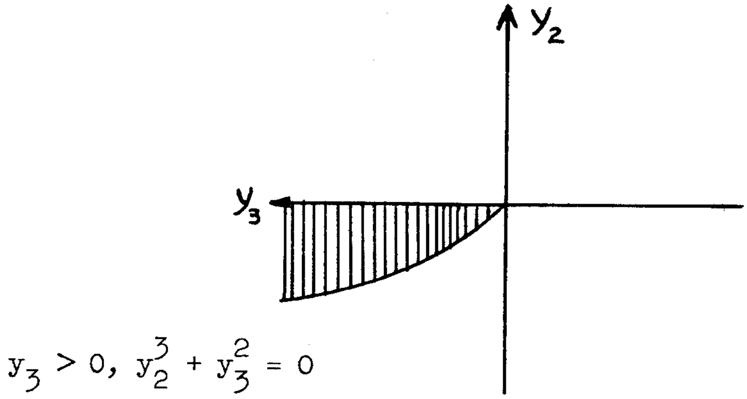


FIGURE 3.1

Thus $R_{31}^P \cap \{(y_2,y_3,y_4)|y_2 > 0,\ y_3 > 0\}$ is the locus of all points (y_2,y_3,y_4) which lie in this quadrant and correspond to points on branch I. All points in the quadrant $y_2 > 0$, $y_3 > 0$, which correspond to points on branch II are spurious. This completes the proof of Theorem 1.

(ii) $\underline{y_2 > 0,\ \ y_3 < 0}$

Theorem 2. The set $R_{31}^P \cap \{(y_2,y_3,y_4)|y_2 > 0, y_3 < 0\}$ is the locus of all points (y_2,y_3,y_4), $y_2 > 0$, $y_3 < 0$, which correspond to points on branch II. All points in the quadrant $y_2 > 0$, $y_3 < 0$, which correspond to points on branch I are spurious.

Proof: We note, first, from (3.11) that Φ is a function of (ζ,ξ) alone. Hence, by Theorem 1,

$$\frac{1-\Phi}{1+\frac{\zeta}{4}\Phi^2} < 0, \qquad \frac{1-\Phi}{1+\frac{\zeta}{4}\Phi^2} - \Phi < 0$$

throughout branch I. Thus, if (y_2,y_3,y_4) is a point in the quadrant $y_2 > 0$, $y_3 < 0$ which corresponds to a point on branch I, it does not satisfy (3.9), (3.10) and cannot belong to R_{31}^{P}. This completes the proof of the second part of Theorem 2.

On the other hand, by Lemma 3,

$$\frac{1-\Phi}{1+\frac{\zeta}{4}\Phi^2} > 0, \qquad \frac{1-\Phi}{1+\frac{\zeta}{4}\Phi^2} - \Phi > 0$$

for all points of branch II with $\zeta > \zeta_0$ for some sufficiently large (but finite) ζ_0.

Such points, then, give rise to curves which lie in R_{31}^{P} in the quadrant $y_2 > 0$, $y_3 < 0$. However, an argument analogous to the one used in the proof of Theorem 1 eliminates the possibility that $\zeta_0 > 0$. This completes the proof of Theorem 2.

B. The Case When $\zeta < 0$

Investigation of this case was not complete when the present chapter was prepared.

Note: The results discussed in the present chapter cover most of the work entailed in the derivation of a closed control law for the system under consideration. These results assure the possibility of obtaining such a control law. However, a certain amount of work remains to be done.

CHAPTER 17

ON THE CONTROLLABLE LINEAR SYSTEM WITH EIGENVALUES $0, 0, \lambda, -\lambda$

On The Linear System With Eigenvalues $0,0,\lambda,-\lambda$ (Preliminaries)

Initial Computations

We consider the system

$$\begin{aligned}
\dot{x}_1 &= \epsilon \\
\dot{x}_2 &= \lambda x_2 + \epsilon \\
\dot{x}_3 &= -\lambda x_3 + \epsilon \, , \quad \lambda > 0 \\
\dot{x}_4 &= x_1
\end{aligned} \tag{1}$$

It is a linear system with eigenvalues $0,0,\lambda,-\lambda$ and one control element. Following our standard procedure we define two sets of auxiliary variables (y_1,y_2,y_3,y_4) and (z_1,z_2,z_3,z_4) as follows:

$$\begin{aligned}
y_1 &= x_1 \\
y_2 &= -1 + e^{-\lambda x_1}(\lambda x_2 + 1) \\
y_3 &= 1 + e^{\lambda x_1}(\lambda x_3 - 1) \\
y_4 &= x_4 - \frac{1}{2} x_1^2
\end{aligned} \tag{2}$$

$$\begin{aligned} z_1 &= -x_1 \\ z_2 &= -1 - e^{+\lambda x_1}(\lambda x_2 - 1) \\ z_3 &= 1 - e^{-\lambda x_1}(\lambda x_3 + 1) \\ z_4 &= -(x_4 + \tfrac{1}{2} x_1^2) \end{aligned} \tag{3}$$

The transformation (2) reduces system (1) to the form $\dot{y}_1 = 1$, $\dot{y}_2 = \dot{y}_3 = \dot{y}_4 = 0$ when $\epsilon = +1$, whereas (3) reduces the system (1) to the form $\dot{z}_1 = 1$, $\dot{z}_2 = \dot{z}_3 = \dot{z}_4 = 0$ when $\epsilon = -1$. The inverse of (2) is given by

$$\begin{aligned} x_1 &= y_1 \\ x_2 &= \frac{1}{\lambda}\,[e^{\lambda y_1}(y_2 + 1) - 1] \\ x_3 &= \frac{1}{\lambda}\,[e^{-\lambda y_1}(y_3 - 1) + 1] \\ x_4 &= y_4 + \tfrac{1}{2} y_1^2 \end{aligned} \tag{4}$$

while the inverse of (3) is

$$
\begin{aligned}
x_1 &= -z_1 \\
x_2 &= \frac{1}{\lambda}\,[-e^{\lambda z_1}(z_2 + 1) + 1] \\
x_3 &= \frac{1}{\lambda}\,[-e^{-\lambda z_1}(z_3 - 1) - 1] \\
x_4 &= -\,(z_4 + \frac{1}{2}\,z_1^2)
\end{aligned}
\tag{5}
$$

Equations (2) and (5) may be used to obtain the transformation from the y's to the z's, namely

$$
\begin{aligned}
y_1 &= -z_1 \\
y_2 &= -1-e^{\lambda z_1}[(z_2 + 1)e^{\lambda z_1} -2] \\
y_3 &= 1-e^{-\lambda z_1}[(z_3 - 1)e^{-\lambda z_1} + 2] \\
y_4 &= -\,(z_4 + z_1^2)
\end{aligned}
\tag{6}
$$

The transformation from the y's to the z's is involutory. This fact follows directly from our general theory (see Chapter 4, Vol. 1, FPR), or it may be checked directly from (3) and (4). The transformation from the z's to the y's is therefore given by

$$
\begin{aligned}
z_1 &= -y_1 \\
z_2 &= -1-e^{\lambda y_1}[(y_2+1)e^{\lambda y_1}-2] \\
z_3 &= +1-e^{-\lambda y_1}[(y_3-1)e^{-\lambda y_1}+2] \\
z_4 &= -(y_4+y_1^2).
\end{aligned}
\tag{7}
$$

The first leaf of the one-dimensional switching curve, denoted by R_{11}, is given by $y_1 < 0$, $y_2 = y_3 = y_4 = 0$. In terms of z these equations become

$$
R_{11}: \begin{cases} z_1 > 0 \\ -1-e^{\lambda z_1}[(z_2+1)e^{\lambda z_1}-2] = 0 \\ +1-e^{-\lambda z_1}[(z_3-1)e^{-\lambda z_1}+2] = 0 \\ -(z_4+z_1^2) = 0 \end{cases}
\tag{8}
$$

We now wish to eliminate z_1 from two of the last three equations in (8). To do this we may use the computations previously carried out for the third order system with eigenvalues $0, \lambda, -\lambda$ [see FPR, Vol. 1, pp. 108-110], since for that system the elimination was effected between two equations which are identical with the first two equations in (8). The effect of this operation is to reduce the equations and inequality defining R_{11} to the form

$$R_{11}:\begin{cases} z_2z_3+ z_2^2z_3+ z_2^2 < 0 \\ e^{\lambda z_1} = (z_2- z_3- z_2z_3)/2z_2 \\ (z_3- z_2+ z_2z_3)^2 + 4z_2z_3= 0 \\ z_4+ \frac{1}{\lambda^2} \log^2(\frac{z_2- z_3- z_2z_3}{2z_2}) = 0 \end{cases} \tag{9}$$

Since $\lambda > 0$ the value of $e^{\lambda z_1}$ on R_{21} is less than its value on R_{11}. It follows that R_{21} is given by

$$R_{21}:\begin{cases} z_2z_3+ z_2^2z_3+ z_2^2 < 0 \\ e^{\lambda z_1} < (z_2- z_3-z_2z_3)/2z_2 \\ (z_3- z_2+ z_2z_3)^2 + 4z_2z_3= 0 \\ z_4+ \frac{1}{\lambda^2} \log^2(\frac{z_2- z_3- z_2z_3}{2z_2}) = 0 \end{cases} \tag{10}$$

The computation of R_{31} requires one final task, namely: we must express the relations (10) in terms of the y's and then eliminate y_1 between the last two equations. To say that this task is formidable is to indulge in understatement. At this stage we have neither succeeded nor given up.

CHAPTER 18

ON A NEW THEORY OF ELIMINATION

On A New Theory Of Elimination

In Chapter 3 of Volume 1, we emphasized the importance of elimination methods for problems of optimal control, and on pp. 42-51 of Vol. 1, we discussed the use of the so-called Weierstrass Preparation Theorem for effecting the required elimination. Unfortunately the practical application of the Weierstrass Preparation Theorem was fraught with considerable difficulty. Not only was it extremely hard to obtain satisfactory expressions for the coefficients in the Weierstrass polynomials but, even assuming the two Weierstrass polynomials of degrees m and n, say, were at hand, the subsequent desired elimination involved, by the dialytic method of Sylvester, the evaluation of a determinant of order $m + n$.

We have now discovered a method of by-passing both the Weierstrass Preparation Theorem and the Sylvester dialytic method. It involves the evaluation of two determinants, each of order only m or n (whichever is the lesser) instead of $m + n$. Nevertheless many difficulties still remain. For one thing the new method rests very extensively on contour integration in the complex plane (as is true also of one method for obtaining the coefficients in the Weierstrass polynomials). Thus a successful application of this new method depends on an expeditious method for carrying out these complex integrations.

In the sequel we give an account of the present status of the new theory, together with a simple example to show its relationship to the Sylvester dialytic method. It should be mentioned, however, that the Sylvester method is applicable to polynomials only, while the new method is applicable to analytic functions.

THEOREM 1. Let $f(z)$ and $g(z)$ be analytic in a region R and let $f(z) \neq 0$ on ∂R. Let the equation $f(z) = 0$ have n roots in R, each root being counted a number of times equal to its multiplicity (so that $n = \frac{1}{2\pi i} \int_{\partial R} \frac{f'(z)}{f(z)} dz$). Then a necessary and sufficient condition that either two roots of the equation $f(z) = 0$ coalesce or that the two equations $f(z) = 0$ and $g(z) = 0$ have a root in common is that

$$\begin{vmatrix} S_0 & S_1 & S_2 & \cdots & S_{n-1} \\ S_1 & S_2 & S_3 & \cdots & S_n \\ S_2 & S_3 & S_4 & \cdots & S_{n+1} \\ \vdots & \vdots & \vdots & & \vdots \\ S_{n-1} & S_n & S_{n-1} & \cdots & S_{2n-2} \end{vmatrix} = 0 \qquad (1)$$

where

$$S_k = \frac{1}{2\pi i} \int_{\partial R} \frac{z^k f'(z) g(z) dz}{f(z)}$$

PROOF: It is known from the theory of analytic functions that

$$S_k = \sum_{i=1}^{n} z_i^k \, g(z_i), \quad k = 0,1,2, \ldots \qquad (2)$$

where $z_1, z_2, \ldots, z_n$ are the n roots of $f(z) = 0$. It is also known that, if any pair of these roots coincide, then the so-called Vandermonde determinant

$$V = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ z_1 & z_2 & z_3 & \cdots & z_n \\ z_1^2 & z_2^2 & z_3^2 & \cdots & z_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ z_1^{n-1} & z_2^{n-1} & z_3^{n-1} & \cdots & z_n^{n-1} \end{vmatrix}$$

must vanish, and conversely. Hence we readily deduce the fact that if any two of the roots $z_1, \ldots, z_n$ of $f(z) = 0$ coincide or if at least one of these numbers is also a root of $g(z) = 0$, then the determinant

$$\Delta = \begin{vmatrix} g(z_1) & g(z_2) & g(z_3) & \cdots & g(z_n) \\ z_1 g(z_1) & z_2 g(z_2) & z_3 g(z_3) & \cdots & \\ z_1^2 g(z_1) & z_2^2 g(z_2) & z_3^2 g(z_3) & \cdots & z_n^2 g(z_n) \\ \vdots & \vdots & \vdots & & \vdots \\ z_1^{n-1} g(z_1) & z_2^{n-1} g(z_2) & z_3^{n-1} g(z_3) & \cdots & z_n^{n-1} g(z_n) \end{vmatrix}$$

must vanish, and conversely. For, of course $\Delta = Vg(z_1)g(z_2)\ldots g(z_n)$. Hence, assuming that at least two of the z_i's coalesce, or that at least one of the $g(z_i)$'s is zero, or both, there must exist n numbers, $\gamma_0, \gamma_1, \gamma_2, \ldots, \gamma_{n-1}$, not all zero, such that

$$\sum_{j=0}^{n-1} \gamma_j z_i^j g(z_i) = 0, \quad i = 1,2,\ldots, n \tag{3}$$

Multiplying (3) by $z_i^\ell (\ell = 0,1,2,\ldots, n-1)$ and summing over i, we obtain

$$\sum_{j=0}^{n-1} \gamma_j \left| \sum_{i=1}^{n} z_i^{j+\ell} g(z_i) \right| = 0, \quad \ell = 0,1,2, \ldots, n-1 \tag{4}$$

It now follows from (2) that

$$\sum_{j=0}^{n-1} \gamma_j S_{j+\ell} = 0, \qquad \ell = 0,1,2, \ldots, n-1 \tag{5}$$

Since the γ's are not all zero, we thus immediately obtain (1) as a necessary condition.

Conversely, if (1) is satisfied there exist n numbers, γ_o, γ_1, γ_2, $\ldots$, γ_{n-1} not all zero, such that (5) is satisfied, whence with the help of (2) we find that (4) is also satisfied. But we can write (4) in the form $\sum_{i=1}^{n} (\sum_{j=0}^{n-1} \gamma_j z_i^j) z_i^\ell g(z_i) = 0$, or better yet in the form

$$\sum_{i=1}^{n} \beta_i z_i^\ell g(z_i) = 0, \qquad \ell = 0,1,2,\ldots, n-1 \tag{6}$$

where

$$\beta_i = \sum_{j=0}^{n-1} \gamma_j z_i^j, \qquad i = 1,2,3, \ldots, n \tag{7}$$

Now,if (Case 1) the β's are not all zero, we see from (6) that the determinant Δ must vanish, whence at least two of the z_i's must coalesce or at least one of the $g(z_i)$'s must vanish, or both. On the other hand, if (Case 2) the β's are all zero, it follows from (7) that $\sum_{j=0}^{n-1} \gamma_j z_i^j = 0$ and since the γ's are not all zero it follows that V must vanish, whence at least two of the z's must coalesce. In either case, we find that (1) is a sufficient condition as stated in the theorem.

THEOREM 2. Suppose that the n roots of the equation $f(z) = 0$ which lie within R, as in the preceding theorem, are distinct and suppose that just one of these roots also satisfies the equation $g(z) = 0$. Then, using the same notation as in the previous theorem, the n homogeneous linear equations,

$$\sum_{j=0}^{n-1} S_{\ell+j}\,\gamma_j = 0, \qquad \ell = 0,1,\ldots, n-1 \tag{8}$$

in the n unknowns $\gamma_0,\gamma_1,\ldots,\gamma_{n-1}$ have a solution for which $\gamma_{n-1} = 1$ and such that the root common to the two equations $f(z) = 0$ and $g(z) = 0$ is equal to

$$\frac{1}{2\pi i}\int_{\partial R} \frac{zf'(z)}{f(z)}\,dz + \gamma_{n-2}$$

PROOF: According to the preceding proof we know that there exist n numbers $\gamma_0,\gamma_1, \ldots,\gamma_{n-1}$ not all zero such that (3) holds and that these γ's satisfy the equations (5). Since just one of the $g(z_i)$'s, say $g(z_n)$, is equal to zero, we can divide each of the first $n-1$ of equations (3) by $g(z_i)$. Thus

$$\sum_{j=0}^{n-1} \gamma_j z_i^j = 0, \quad i = 1,2,\ldots,n-1$$

This means that the (n-1) roots, distinct from z_n, satisfy the algebraic equation

$$\gamma_{n-1} z^{n-1} + \gamma_{n-2} z^{n-2} + \ldots + \gamma_o = 0$$

Moreover, $\gamma_{n-1} \neq 0$, since otherwise the degree of this equation would be less than (n-1) while it still would admit (n-1) distinct roots. Since equations (3) as well as (5) or the equivalent equations (8), are homogeneous, we may choose $\gamma_{n-1} = 1$. Hence the sum of the (n-1) roots distinct from z_n is equal to $-\gamma_{n-2}$. Since the sum of <u>all</u> n roots of the equation $f(z) = 0$ is known to be equal to

$$\frac{1}{2\pi i} \int_{\partial R} \frac{zf'(z)}{f(z)} dz,$$

we find by subtraction that

$$z_n = \frac{1}{2\pi i} \int_{\partial R} \frac{zf'(z)}{f(z)} dz + \gamma_{n-2}$$

as we wished to prove.

From Theorem 1 we, of course, have a quantity Φ, namely the determinant of (1), which vanishes whenever f and g have common zeros in R. Unfortunately, however, it also vanishes under circumstances when f and g do not have a common zero, namely whenever f has multiple zeros. If $f(z)$ and $g(z)$ depend analytically upon another complex variable ζ (or upon several such variables), then Φ also depends analytically upon ζ, and it is, then possible, in general, by way of the theory of removable singularities, to define a function $\Psi(\zeta)$ which vanishes if, and only if, f and g have at least one common zero, this common zero being, of course, a function of ζ. To show how this comes about we first prove the following:

LEMMA 1. Let S denote the matrix

$$\begin{bmatrix} s_o & s_1 & s_2 & \cdots & s_{n-1} \\ s_1 & s_2 & s_3 & \cdots & s_n \\ s_2 & s_3 & s_4 & \cdots & s_{n+1} \\ \vdots & \vdots & \vdots & & \vdots \\ s_{n-1} & s_n & s_{n+1} & \cdots & s_{2n-2} \end{bmatrix}$$

where

$$S_k = \frac{1}{2\pi i} \int_{\partial R} \frac{z^k g(z) f'(z)}{f(z)} dz, \quad k = 0,1,2,\ldots$$

and let M denote the Vandermonde matrix

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ z_1 & z_2 & z_3 & \cdots & z_n \\ z_1^2 & z_2^2 & z_3^2 & \cdots & z_n^2 \\ \vdots & \vdots & \vdots & & \vdots \\ z_1^{n-1} & z_2^{n-1} & z_3^{n-1} & \cdots & z_n^{n-1} \end{bmatrix}$$

Then

$$\det S = g(z_1)g(z_2)\ldots g(z_n)[\det M]^2 \tag{9}$$

or, in terms of other previously introduced notation, whereby Φ = det S and V = det M,

$$\Phi = g(z_1)g(z_2)\ldots g(z_n)V^2 . \tag{10}$$

PROOF: From the definition of Δ given in the proof of Theorem 1 we have

$$g(z_1)g(z_2)\ldots g(z_n)V = \Delta \tag{11}$$

Hence, it is sufficient to establish the matrix equality

$$S = (\text{matrix of } \Delta)M' \tag{12}$$

where M' is the transpose of the Vandermonde matrix M. For, if this matrix equation were established we would have

$$\det S = \Delta \det M' = \Delta \det M = \Delta \cdot V \tag{13}$$

whereas, from (11) we know that

$$\Delta = g(z_1)g(z_2)\ldots g(z_n)V,$$

so that, upon inserting this value of Δ into (13), we obtain

$$\det S = g(z_1)g(z_2)\ldots g(z_n)V\cdot V$$

which is equivalent to (9) or (10).

To establish (12) note that the $(p + 1)$th row of the matrix of Δ contains the elements $z_1^p g(z_1)$, $z_2^p g(z_2), \ldots, z_n^p g(z_n)$, whereas the $(q + 1)$th column of M', (which is the same at the $(q + 1)$th row of M) contains the elements $z_1^q, z_2^q, \ldots, z_n^q$. Hence, by the rule for forming matrix products, the element in the $(p + 1)$th row and (q + 1)th column of (matrix of $\triangle$) M' must be $\sum_{i=1}^{n} z_i^{p+q} g(z_i)$, which, by (2), is the same as S_{p+q}, the element in the (p + 1)th row and (q + 1)th column of S, as we wished to prove.

Having completed the proof of Lemma 1, we now introduce the hypothesis that f and g depend analytically on ζ as long as ζ belongs to a specified domain D. Then the solution of the equation $f(z,\zeta) = 0$ for z in terms of ζ is an n-valued function of ζ, analytic except possibly for branch points. We assume that all n branches lie in R as long as $\zeta \in D$. However, any analytic symmetric function of these n branches, denoted by $z_1(\zeta)$, $z_2(\zeta)$, ..., $z_n(\zeta)$, must be analytic without even branch points, in its dependence upon $\zeta \in D$. In particular, this is true of the product

$$g(z_1(\zeta),\zeta)g(z_2(\zeta),\zeta)\dots g(z_n(\zeta),\zeta)$$

and also of the square of the Vandermonde determinant,

$$V(\zeta) = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ z_1(\zeta) & z_2(\zeta) & z_3(\zeta) & \cdots & z_n(\zeta) \\ z_1(\zeta)^2 & z_2(\zeta)^2 & z_3(\zeta)^2 & \cdots & z_n(\zeta)^2 \\ \vdots & \vdots & \vdots & & \vdots \\ z_1(\zeta)^{n-1} & z_2(\zeta)^{n-1} & z_3(\zeta)^{n-1} & \cdots & z_n(\zeta)^{n-1} \end{vmatrix} \qquad (14)$$

If we assume that $V(\zeta)^2$ does not vanish identically in ζ, then we know from the theory of analytic functions that it vanishes only at isolated points. The quantities

$$S_k = \frac{1}{2\pi i} \int_{\partial R} \frac{z^k g(z,\zeta) f'(z,\zeta)}{f(z,\zeta)}\, dz, \quad k = 0,1,2,\ldots$$

are also clearly analytic functions of ζ and hence so is $\Phi = \det S$. It follows, at once, that the function

$$\Psi(\zeta) = \frac{\Phi(\zeta)}{V(\zeta)^2}$$

is also analytic in D except possibly at points where $V(\zeta)$ vanishes where Ψ is not even defined. But from (10) it is clear that except at these isolated points where it is not defined

$$\Psi(\zeta) = g(z_1(\zeta),\zeta)g(z_2(\zeta),\zeta)\cdots g(z_n(\zeta),\zeta) \tag{15}$$

Hence (15) can be used to define Ψ also at its isolated singular points, and the resulting $\Psi(\zeta)$ is analytic throughout D. Moreover it is obvious from (15) that $\Psi(\zeta)$ vanishes if and only if the value of ζ is such that one (or more) of the quantities $g(z_1(\zeta),\zeta)$, $g(z_2(\zeta),\zeta),\ldots, g(z_n(\zeta),\zeta)$ is zero. That is, $\Psi(\zeta) = 0$ if and only if ζ takes on a value such that the equations

$$f(z,\zeta) = 0 \quad \text{and} \quad g(z,\zeta) = 0 \tag{16}$$

have a common solution for z.

It may be added that, instead of using (14), we can use the fact that

$$V(\zeta)^2 = \begin{vmatrix} \sigma_0 & \sigma_1 & \sigma_2 & \cdots & \sigma_{n-1} \\ \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_n \\ \sigma_2 & \sigma_3 & \sigma_4 & \cdots & \sigma_{n+1} \\ \vdots & \vdots & \vdots & & \vdots \\ \sigma_{n-1} & \sigma_n & \sigma_{n+1} & \cdots & \sigma_{2n-2} \end{vmatrix} \tag{17}$$

where

$$\sigma_k = \frac{1}{2\pi i} \int_{\partial R} \frac{z^k f'(z,\zeta)}{f(z,\zeta)}\, dz, \quad f'(z,\zeta) = \frac{\partial(z,\zeta)}{\partial z}$$

This follows from Lemma 1 in the special case $g(z,\zeta) = 1$.

We summarize these results in the following:

THEOREM 3. Let $f(z,\zeta)$ and $g(z,\zeta)$ be analytic in $z \in R$ and $\zeta \in D$ and let f for each $\zeta \in D$ have n zeros located in R but suppose that $f(z,\zeta) \neq 0$ for $z \in \partial R$ and $\zeta \in D$. Assume that the determinant in formula (17) does not vanish identically in D. Then there exists a function $\Psi(\zeta)$ analytic in D which vanishes at those points ζ of D (and only at those points) for which the equations (16) have a common solution. Moreover at points where $V(\zeta) \neq 0$, $\Psi(\zeta) = \Phi(\zeta)/V(\zeta)^2$ where $\Phi = \det S$ and $V(\zeta)$ is given by (17).

In applying these results to cases where f and g are polynomials and the region R is a circle of sufficiently large radius centered at the origin, it is necessary to evaluate integrals of the form

$$\frac{1}{2\pi i} \int_{\partial R} \frac{P(z)}{Q(z)}\, dz = \lim_{\beta \to \infty} \frac{1}{2\pi i} \int_{C(\beta)} \frac{P(z)}{Q(z)}\, dz$$

where $C(\beta)$ represents the circle with radius β and center at the origin, and where P and Q are polynomials in z. If the degree of P is not less than the degree of Q, we find by division algorithm that

$$\frac{P(z)}{Q(z)} = A(z) + \frac{B(z)}{Q(z)}$$

where A and B are polynomials and the degree of B is less than the degree of Q. Since A(z) is analytic we know that

$$\int_{C(\beta)} A(z)dz = 0$$

Hence

$$\frac{1}{2\pi i} \int_{\partial R} \frac{P(z)}{Q(z)} dz = \lim_{\beta \to \infty} \frac{1}{2\pi i} \int_{C(\beta)} \frac{B(z)}{Q(z)} dz$$

Let $Q(z) = q_o z^{\mu} + q_1 z^{\mu-1} + qz^{\mu-2} + \ldots,\quad q_o \neq 0,$ and $B(z) = b_o z^{\mu-1} + b_1 z^{\mu-2} + b_2 z^{\mu-3} + \ldots$

then $$\frac{1}{2\pi i} \int_{\partial R} \frac{P(z)}{Q(z)} dz = \lim_{\beta \to \infty} \frac{1}{2\pi i} \int_{C(\beta)} \frac{1}{z}\left[\frac{b_o + b_1 z^{-1} + b_2 z^{-2} + \ldots}{q_o + q_1 z^{-1} + q_2 z^{-2} + \ldots}\right]dz =$$

$$\lim_{\beta \to \infty} \frac{1}{2\pi i} \int_0^{2\pi} \left[\frac{b_o + b_1\beta^{-1}e^{-i\theta} + b_2\beta^{-2}e^{-2i\theta} + \ldots}{q_o + q_1\beta^{-1}e^{-i\theta} + q_2\beta^{-2}e^{-2i\theta} + \ldots}\right]d\theta,$$

and since it is easy to establish the uniform convergence of the last written integrand, as $\beta \to \infty$, to b_o/q_o, we reach the result that

$$\frac{1}{2\pi i} \int_{\partial R} \frac{P(z)}{Q(z)}\, dz = \frac{b_o}{q_o} \tag{18}$$

Of course, if the degree of $B(z)$ is less than $\mu-1$, we take $b_o = 0$. But we always have $q_o \neq 0$ by definition of μ as the degree of $Q(z)$.

We are now in a position to apply Theorems 1, 2, and 3 to the case where

$$f(z) = z^2 + bz + c \quad \text{and} \quad g(z) = z^2 + pz + q$$

and where R is any region large enough to contain both roots of the equation $f(z) = 0$. In applying Theorem 3 we may take p and q to be constants and also either b or c. The other one may be taken as ζ and the region D may be regarded as the entire complex plane.

We must first calculate S_o, S_1, and S_2 as defined in Theorem 1. For instance,

$S_2 = \frac{1}{2\pi i} \int_{\partial R} \frac{z^2(2z + b)(z^2 + pz + q)}{z^2 + bz + c}\, dz$ and by the division algorithm we have

$$\frac{z^2(2z + b)(z^2 + p + pz + q)}{z^2 + bz + c} = 2z^3 + (-b + 2p)z^2 + (b^2 - pb + 2q - 2c)z +$$

$$(-b^3 + pb^2 - qb + 3cb - 2pc) + \frac{(b^4 - pb^3 + (-4c + q)b^2 + 3pcb + (2c^2 - 2qc))z + \lambda}{z^2 + bz + c}$$

where λ is a quantity whose value is irrelevant. Hence, we find by the method explained above that

$$S_2 = b^4 - pb^3 + (-4c + q)b^2 + 3pcb + (2c^2 - 2qc) \qquad (19)$$

Similarly we find that

$$S_1 = -b^3 + pb^2 - qb + 3cb - 2pc \qquad (20)$$

and that

$$S_o = b^2 - pb + 2q - 2c \qquad (21)$$

(The fact that S_1 and S_o are certain coefficients in the quotient is not entirely accidental as the reader will soon discover if he carries out in detail the calculations of S_o, S_1, S_2 by use of (18)).

The quantities σ_0, σ_1, and σ_2 are obtained in the same way, but with far less computation: thus

$$\sigma_2 = \frac{1}{2\pi i} \int_{\partial R} \frac{2z^3 + bz^2}{z^2 + bz + c} dz = b^2 - 2c \tag{22}$$

$$\sigma_1 = \frac{1}{2\pi i} \int_{\partial R} \frac{2z^2 + bz}{z^2 + bz + c} dz = -b \tag{23}$$

$$\sigma_0 = \frac{1}{2\pi i} \int_{\partial R} \frac{2z + b}{z^2 + bz + c} dz = 2 \tag{24}$$

It now follows from (19)-(24), from the definition of $\Phi = \det S$, and from (17) that

$$\Phi = qb^4 + (-pc - pq)b^3 + (c^2 - 6qc + p^2c + q^2)b^2 + (4pc^2 + 4pqc)b + (-4c^3 + 8c^2q - 4cq^2 - 4p^2c^2)$$

and

$$V(\zeta)^2 = b^2 - 4c.$$

This last expression is, of course, precisely the discriminant of $z^2 + bz + c$, as it should be, and, for this reason it was not really necessary to carry out the calculations indicated in (22)-(24).

The essence of Theorem 3 is to the effect that Φ is exactly divisible by $V(\zeta)^2$. This fact is readily verifiable in the present example. In fact it is found from the above expression for Φ that

$$\Phi = (b^2-4c)(-pcb-pqb + c^2+ p^2c + q^2+ qb^2-2qc)$$

Thus, the quantity Ψ of Theorem 3 turns out, in this example, to be

$$\Psi = -pcb-pqb + c^2+ p^2c + q^2+ qb^2-2qc$$

This turns out to be exactly the Sylvester eliminant.

$$\Psi = \begin{vmatrix} 1 & p & q & 0 \\ 0 & 1 & p & q \\ 1 & b & c & 0 \\ 0 & 1 & b & c \end{vmatrix}$$

In order to illustrate Theorem 2, it is necessary to find the quantities γ_o, γ_1, which according to (5) are given in this case $n = 2$, by the equations

$$S_o\gamma_o + S_1\gamma_1 = 0$$

$$S_1\gamma_o + S_2\gamma_1 = 0$$

We are interested in the case when these two equations have a simultaneous non-trivial solution and, in fact, according to Theorem 2 we look for a solution in which $\gamma_1 = 1$. Thus we find that $\gamma_o = -S_1/S_o$ and the formula of Theorem 2 for the root r common to the two equations $f(z) = 0$ yields, and $g(z) = 0$ yields

$$r = \frac{1}{2\pi i} \int_{\partial R} \frac{zf'(z)}{f(z)} dz + \gamma_o = \sigma_1 - \frac{S_1}{S_o} = \frac{S_o\sigma_1 - S_1}{S_o}$$

Hence, using (20), (21) and (23), we get

$$r = \frac{2pc-bc + bq}{b^2 - pb + 2q-2c}$$

In view of the fact that we are dealing with the case where $\Psi = 0$, i.e.,

$$qb^2 - 2qc - pqb - pbc + q^2 + c^2 + p^2c = 0$$

we find the following equivalent but simpler expression for r:

$$r = \frac{c-q}{p-b}$$

Either of these expressions for the common root of the two equations is, of course, in this example also obtainable by the dialytic method of Sylvester.

CHAPTER 19

TIME OPTIMAL CONTROL

SUBJECT TO PHASE COORDINATE CONSTRAINTS

1. Statement Of The General Problem

Consider the system

$$\dot{x} = f(x) + au \qquad (1.1)$$

where x is an n-vector representing the system's state, $f(x)$ is an n-vector function of x, a is a constant n-vector and $u = u(t)$, the control parameter, is a scalar function to be more properly described below. We assume that f is of class C^2 in some region G containing the origin and that $f(0) = 0$. Moreover, we assume that the origin is an isolated zero of f. The function u is restricted to the class U of all real valued piecewise continuous functions on the real line whose range is contained in the closed interval $[-1, 1]$. The space of x is denoted by X.

A point x_o in G is said to be controllable if there exists a function $u \in U$ which steers the system from x_o to the origin in finite time. The set of all controllable points in G is called the controllable region in X and denoted by R.

Let $N \subset X$ be a given closed set in phase space. The set N will be called the set of constraints. Consider the set N_1 of all those points $x \in R$ for which there exists a control function

$u_x(t) \in U$ which steers the point x to the origin without ever leaving the set N. Clearly $N_1 \subset N \subset R$, for if $x \in N_1$ then x may be steered into the origin, whence $x \in R$, and on the other hand x cannot be outside N without violating the condition that x be steered into the origin without leaving N. If $x_o \in N_1$ then there exists at least one control function $u_{x_o}(t)$ which steers x_o to the origin in finite time without ever leaving the set N. However, the function $u_{x_o}(t)$ is not necessarily unique. In fact, there may exist infinitely many distinct control functions each of which steers x_o to the origin without ever leaving the set N. Denote the set of all these control functions by $U_{x_o}(N)$. Our problem may now be stated as follows: for a given point $x_o \in N_1$ find that function (or those functions) in $U_{x_o}(N)$ which steer the point x_o to the origin in minimum time $T = T(x_o,N)$. It may, of course, happen that this problem, as formulated above, is too severe. It may not be generally possible to find an optimal control for every point x_o in the set N_1. Although every point in N_1 is controllable within N, the search for an optimal control may have to be restricted to a set smaller than N_1.

2. On The Notion Of Controllability.

Let $\Gamma_x(0)$ be the class of all (unconstrained) admissible

[redacted]

trajectories which connect a fixed point $x \in R$ to the origin. We shall restrict our attention to systems (1.1) which have the property that at most one of the members of $\Gamma_x(0)$ satisfies the maximum principle (all $x \in R$). In other words, we assume that if there is a solution satisfying the necessary conditions for optimality embodied in the maximum principle, then this solution is unique.

We shall say that a set $K \subset R$ is controllable within a set $M \subset X$ if for every point $x \in K$ there exists an admissible control $u_x(t)$ which steers x to the origin in finite time without ever leaving the set M. Using this formulation the set N_1 may be defined as the maximal subset which is controllable within N. It is easy to see that R is controllable within itself. For if $x \in R$ then there exists an admissible control which steers x to the origin. If y is any intermediate point on an admissible trajectory which connects x to the origin, then y too is controllable. It follows that every admissible trajectory is contained in R, whence R is controllable within itself. Thus, if $N = R$, the problem of time-optimal control subject to constraints is identical with the unconstrained problem.

Suppose N is a proper subset of R. Let $x_o \in N_1$ and let $u_{x_o}(t)$ be an admissible control which is time-optimal relative to the class $U_{x_o}(N)$. Let Γ be the trajectory corresponding to $u_{x_o}(t)$ which connects x_o to the origin. It follows from the definition of $U_{x_o}(N)$ that $\Gamma \subset N$. If Γ does not intersect the boundary of N then there is a whole neighborhood of Γ which lies in the interior of N. Hence Γ must satisfy the maximum principle throughout its length. It therefore follows from our assumption concerning the system (1) that Γ is optimal relative to the whole class U. Thus, if Γ does not intersect the boundary of N it is identical with the optimal trajectory of the unconstrained problem. Otherwise, Γ is composed of arcs which lie alternately in the interior of N and on its boundary. Following standard notation we shall denote the boundary of N by ∂N.

Let N_2 be that subset of N_1 which has the property that every one of its points has an <u>optimal</u> control which steers it to the origin (within N). In other words, N_2 is the set of all those points $x_o \in N_1$ for which there exists a control which is optimal relative to the class $U_{x_o}(N)$. A point $x \in N_2$ will be said to be <u>strongly controllable</u> . Clearly $N_2 \subset N_1 \subset N$. Examples in which $N_1 \neq N$ will be given in the next section. However, we have not yet found an example of a case in which $N_2 \neq N_1$, nor have we succeeded in proving that N_2 must equal N_1. As assertion to the effect that

$N_1 = N_2$ (under certain reasonable assumptions) would be important inasmuch as it would establish the existence of a solution to the problem of time-optimal control with constraints throughout the set N_1. On the other hand, if N_2 is not necessarily equal to N_1, there would be points in N, which are controllable within N, but are not strongly controllable there (they may be strongly controllable without constraints). As stated above, this question is still open.

Before proceeding further with a detailed discussion of results obtained by us, it seems appropriate to relate the problem at hand to some rather far reaching theorems in the calculus of variations which are found in the literature. The most appropriate treatment for present purposes (especially as regards the first problem) is to be found in Chapter 6 of "The Mathematical Theory of Optimal Processes" by Pontryagin, Boltyansky, Gamkrelidze, and Mischenko (translated by Trirogoff).

The problem considered in that chapter is concerned with the system,

$$\frac{dx}{dt} = f(x, u)$$

where x is a point in a closed region B of n-dimensional space and

u is a point in a closed region of r-dimensional space. Given two points x_o and x_1 in B, one considers the class C of all functions $u(t)$, whose values lie in U and which are defined on some interval $t_o \leqq t \leqq t_1$, such that there exists a solution $x(t)$ of the above system having x_o and x_1 as end points and everywhere contained in B. That is $x(t_o) = x_o$, $x(t_1) = x_1$, and, $x(t) \in B$ for each t on the interval $t_o \leqq t \leqq t_1$. The problem, then, is to choose out of this class C, a particular $u(t)$, which minimizes a given integral of the form,

$$\int_{t_o}^{t_1} f^o(x(t), u(t))dt.$$

The r-vector functions $u(t)$ may assume values on the boundary of U and the n-vector functions $x(t)$ may assume values on the boundary of B. An arc of such an optimal trajectory which lies entirely in the interior of B, except for its end points (which may lie on the boundary of B), must satisfy the Pontryagin maximum principle. If, however, it lies entirely on the boundary of B it must still satisfy a modified maximum principle of lower dimensionality, and there is also a so-called jump condition which must be satisfied at the juncture of two such arcs of either kind.

For certain kinds of problems of particular importance, the maximum principle implies bang-bang control. Evidently, then, the theory of bang-bang control is going to continue to play an important role in the constrained problem.

3. Contributions To The General Theory

We return to the system

$$\dot{x} = f(x) + au \tag{3.1}$$

where x is an n-vector representing the system's state, $f(x)$ is an n-vector function of x, a is a constant n-vector and $u = u(t)$, the control parameter, is a scalar function. We assume that f is of class C^2 in some region containing the origin and that $f(0) = 0$. Moreover, we assume that the origin is an isolated zero of f. The function u is restricted to the class U of all real valued piecewise continuous functions on the real line whose range is contained in the closed interval $[-1,1]$. System (1) is assumed to be controllable in a certain neighborhood of the origin. The space of x is denoted by X.

Let R be the controllable region in X, namely the set of all points which can be steered into the origin in finite time. If x is in R then there exists a control function $u_x(t)$ in U

which steers x into the origin in finite time $T(x, u_x)$. The assumption that there is a neighborhood of the origin which is controllable says that R contains an open set G which contains the origin.

<u>THEOREM 1</u>. If R contains an open set G which contains the origin then R is open.

<u>PROOF:</u> Let x be an arbitrary point in R. Then there exists a control function $u_x(t)$ which steers x into the origin in finite time. Let $S(\delta)$ denote an open sphere of radius δ and center at the origin. Choose δ sufficiently small so that $S(\delta) \subset G$. Since $u_x(t)$ steers x into the origin it must steer it into $S(\frac{\delta}{3})$. Let $\varphi = \varphi(t,x,u)$ denote the solution of system (3.1), corresponding to the control u, which passes through the point x at time $t = 0$. Then there exists a time $t^* > 0$ such that $\varphi(t^*,x,u_x) \in S(\frac{\delta}{3})$. Let p be the point $\varphi(t^*,x,u_x)$ and let $S_p(\frac{\delta}{3})$ be the open sphere of radius $\frac{\delta}{3}$ and center p. Given p and δ there clearly exists a neighborhood N_x of x having the property that if y is any point in N_x then

$$\|\varphi(t,y,u_x) - \varphi(t,x,u_x)\| < \frac{\delta}{3}$$

for all $0 \leqq t \leqq t^*$. (Here $\|p\|$ denotes the norm of p). In particular, this inequality holds for $t = t^*$, whence

$$\varphi(t^*,y,u_x) \in S_p(\tfrac{\delta}{3}) \subset S(\delta) \subset R.$$

Thus the trajectory through y with control $u_x(t)$ intersects R and therefore $y \in R$. Hence $N_x \subset R$ and R is open. This completes the proof.

Let N be a closed bounded (hence compact) subset of R which contains the origin. Following § 1 we denote by N_1 the subset of N which consists of all points which are controllable within N. We are interested in the properties of the set N_1.

PROPOSITION 1. N_1 is not empty.

PROOF: $0 \in N_1$.

PROPOSITION 2. N_1 is the maximal subset of N which is controllable within itself.

PROOF: Let $\{S_\alpha\}$ be the collection of all subsets of N each of which has the property that it is controllable within itself. Let

$$S = \bigcup_\alpha S_\alpha .$$

Clearly $S \subset N$. If $x \in S$, then $x \in S_{\alpha^*}$ for some α^*. Since S_{α^*}

is controllable within itself, x is controllable within $S_{\alpha*}$ and therefore within S. It follows that S is controllable within itself. If K is any subset of N which is controllable within itself then, by definition of S, $K \subset S$. Hence S is the maximal subset of N which is controllable within itself. We shall show that $S = N_1$.

S is controllable within itself and $S \subset N$. Hence, by definition of N_1, $S \subset N_1$. Conversely, let $x \in N_1$. Then there exists a control $u_x(t)$ which steers x to the origin in finite time $T(x,u_x)$ within the set N. The arc $\{\varphi(t,x,u_x) | 0 \leq t \leq T(x,u_x)\}$ is a subset of N and is clearly controllable within itself. Hence it is contained in S and, in particular, $\varphi(0,x,u_x) = x$ is in S. Therefore, $N_1 \subset S$. This completes the proof of Proposition 2.

The set N_1 may actually reduce to a single point, namely the origin. This happens, for example, in the system $\dot{x}_1 = \epsilon$, $\dot{x}_2 = x_1$ for the set N consisting of all points lying on the x_2-axis between $x_2 = -1$ and $x_2 = +1$.

We shall assume henceforth that the set N contains an open set which contains the origin. It is not clear to us at this point whether this implies that the set N_1 has the same property.

<u>LEMMA 1</u>. Let $\delta > 0$ be sufficiently small so that $S(\delta) \subset G$. Let C be a compact subset of R. Then there exists a time $T(C,\delta)$ such that every point in C may be steered into $S(\delta)$ in time $T(C,\delta)$ or less.

<u>PROOF:</u> Let $x \in C$. Then there exists a control $u_x(t) \in U$ which steers x into the origin in time $T(x,u_x)$. Hence, there exists an open neighborhood N_x of x such that

$$\varphi(T(x,u_x),y,u_x) \in S(\tfrac{\delta}{2})$$

for all $y \in N_x$. The collection of neighborhoods $\{N_x | x \in C\}$ form an open covering of C from which we may select a finite subcovering $\{N_{x_1},\dots, N_{x_r}\}$. Pick

$$T(C,\delta) = \max_{i = 1,\dots, r} T(x_i,u_{x_i}).$$

This completes the proof of Lemma 1.

<u>LEMMA 2.</u> For any set F, let F^o denote the interior of F. Suppose there exists an open subset G^* of N which contains the origin and is controllable within itself. Suppose, furthermore, that x is controllable within N^o. Then $x \in N_1^o$.

PROOF: If x is controllable within N^o then $x \in N_1$. Moreover, there exists a control $u_x(t) \in U$ which steers x to the origin in time $T(x,u_x)$ in such a way that $\varphi(t,x,u_x) \in N^o$ for all $0 \leq t \leq T(x,u_x)$. Let $\Gamma = \Gamma(x; 0,T(x,u_x))$ denote the arc of the trajectory of $\varphi(t,x,u_x)$ corresponding to the time interval $[0,T(x,u_x)]$. Γ is an arc in the topological sense of this word, hence it is compact. The boundary ∂N of the set N is also compact and $\partial N \cap \Gamma = 0$. Hence the distance between Γ and ∂N is positive, say $\eta > 0$. Let $G^* \subset N$ be the given open neighborhood of the origin which is controllable within itself. For any fixed positive integer r there exists a neighborhood N_x of x having the property that

$$\|\varphi(t,y,u_x) - \varphi(t,x,u_x)\| < \frac{\eta}{2^r}$$

for all $y \in N_x$ and all $0 \leq t \leq T(x,u_x)$. Choose r large enough so that $S(\frac{\eta}{2^r}) \subset G^*$. Thus $\varphi(t,y,u_x)|0 \leq t \leq T(x,u_x)$ is contained in a tubular neighborhood of Γ which does not intersect ∂N, and $\varphi(T(x,u_x),y,u_x) \in G^*$. Hence y is controllable within N^o and therefore $N_x \subset N_1$. It follows that $x \in N_1^o$. This completes the proof of Lemma 2.

COROLLARY 1. Under the assumption of Lemma 2, if $x \in \partial N_1 \cap N_1$ then any trajectory which steers x to the origin must meet ∂N.

PROOF: If $x \in N_1$ and there exists a control function which steers x to the origin without meeting ∂N, then x is controllable within N^o. But then $x \in N_1^o$, by Lemma 2. This completes the proof.

4. A Two-Dimensional Example

Consider the system

$$\dot{x}_1 = \epsilon, \quad \dot{x}_2 = x_1, \quad |\epsilon| \le 1 \tag{4.1}$$

The controllable region of system (4.1) is the whole plane. Let N, the set of constraints, be a disc of radius R with the center at the origin. Our problem is to find the sets N_1 and N_2 and to develop a time-optimal control law for points in N_2 subject to the constraints represented by the set N.

We know from the theorem mentioned in § 2 [L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, E. F. Mishchenko, THE MATHEMATICAL THEORY OF OPTIMAL PROCESSES, p. 311], that any portion of an optimal trajectory lying in the interior of the set N must satisfy the maximum principle. Hence, such portion of such trajectory will have to be Bang-Bang.

The switching curve of system (4.1) is well known from our previous investigations. Its leaves are given by

R_{11}: $x_2 - \frac{1}{2} x_1^2$, $x_1 < 0$

$$R_{12}: \; x_2 + \frac{1}{2} x_1^2 , \quad x_1 > 0 \tag{4.2}$$

A closed form optimal control law for the unconstrained system was found in Chapter 13 to be

$$\epsilon = -\text{sgn}[x_2 + \frac{1}{2} (\text{sgn } x_1) x_1^2] \tag{4.3}$$

Let $r = (x_1^2 + x_2^2)^{\frac{1}{2}}$. Then $\dot{r} = (x_1/r)(\epsilon + x_2)$. Using the value of ϵ as given in (4.3) we find that $\dot{r}$ is positive throughout the shaded region of Figure 4.1 and negative otherwise. We shall distinguish among the following cases:

(i) $R \leqq 1$

(ii) $1 < R < \sqrt{3}$

(iii) $R \geqq \sqrt{3}$

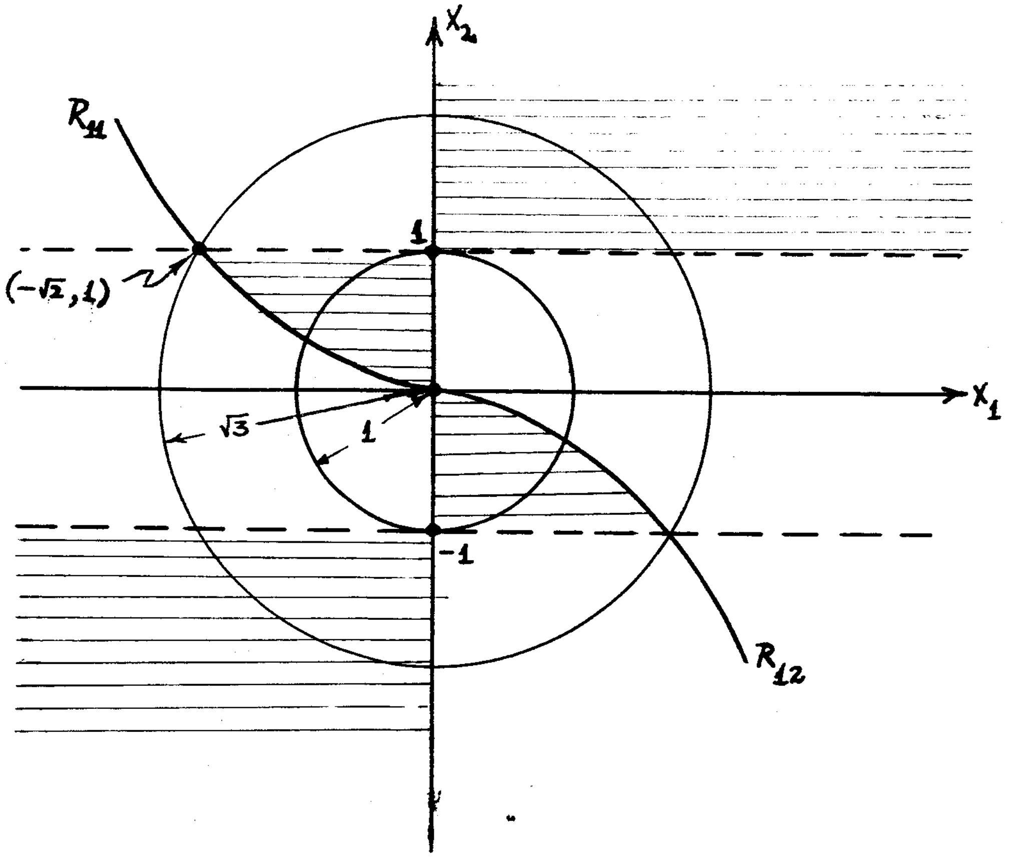


FIGURE 4.1

(i) The Case Where $R \leqq 1$

For points lying. above the switching curve the value of ϵ is -1. For such points the value of $\dot{r}$ is negative in the first and fourth quadrants. Therefore any trajectory starting within the fourth or first quadrant above the switching curve (and, of course, within the disc N, of radius R) cannot leave N either in the fourth or first quadrant. Such trajectory must therefore meet the positive half of the x_2-axis at a point whose x_2-coordinate satisfies $0 < x_2 \leqq R$. However, once the trajectory crosses the x_2-axis its distance from the origin begins to increase. There are two possibilities: the continued trajectory may meet the leaf R_{11} before meeting the boundary of N, in which case the complete trajectory lies in the interior of N and is therefore identical with the unconstrained case, or it may meet the boundary of N before meeting R_{11}. In the latter case, part of the optimal trajectory, if there is one, must lie on the boundary of the disc.

Now it is easy to see that an arc of a trajectory of system (4.1) will lie on a circle with center at the origin if and only if $\epsilon = -x_2$. Since $R \leqq 1$ and $|x_2| \leqq R$ for every point on the boundary of N, the control $\epsilon = -x_2$ is admissible. Since it is unique, it is also optimal. Once the curve R_{11} is reached, either in the interior of N or on its boundary, control is switched to $\epsilon = +1$ and the system is steered into the origin on R_{11}. We conclude, therefore, that for any point P in N, lying above the switching curve (of the unconstrained problem), there exists a unique control which steers it time-optimally to the origin within the set N.

The situation below the switching curve is completely analogous, except that the value assigned to ϵ in the interior of N is reversed. However, the value assigned to ϵ on the boundary of N

remains the same, namely, $\epsilon = -x_2$.

Thus, the case when $R \leq 1$, provides us with an example in which $N = N_1 = N_2$. It is for that reason that a complete solution to the problem is possible: every point in N is controllable within N and moreover, every point within N has a (unique) optimal control within N.

Examples of optimally controlled trajectories within N are given in Figure 4.2

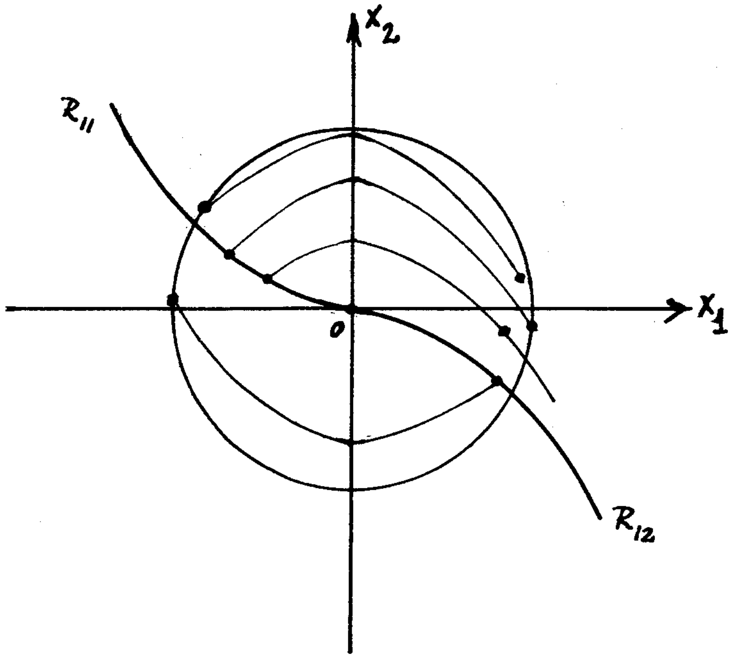


FIGURE 4.2

We now proceed to develop a closed-form time-optimal control law subject to the constraint embodied in the set $N(R \leq 1)$.

Let ϵ be as in (4.3). We shall write ϵ^* for the optimal control law of the constrained system. We shall show that one form of such control is given by

$$\epsilon^* = \frac{1}{2}(1 + \mathrm{sgn}[R-r])\epsilon + \frac{1}{8}(1-\mathrm{sgn}[R-r])(1-\mathrm{sgn}\ x_1)(\{1 + \epsilon\ \mathrm{sgn}[R-r]\}\{-x_2\}$$

$$+ \{1-\epsilon\ \mathrm{sgn}[R-r]\}\epsilon) + \frac{1}{8}(1-\mathrm{sgn}[R-r])(1 + \mathrm{sgn}\ x_1)(\{1-\epsilon\ \mathrm{sgn}[R-r]\}\{-x_2\}$$

$$+ \{1 + \epsilon\ \mathrm{sgn}[R-r]\}\epsilon) \qquad (4.4)$$

We first remark that ϵ^* requires slight overshoots beyond the circle of radius R. This, however, does not create any difficulty. If it were necessary to keep strictly within the disc N, one would simply replace R in (4.4) by a quantity R' which is slightly smaller than R. This would assure control strictly within N.

The function ϵ^* is given in terms of three summands. The first of these vanishes outside the circle of radius R, whereas the last two vanish in its interior. Thus, in the interior of N we have,

$$\epsilon^* = \frac{1}{2}(1 + \mathrm{sgn}[R-r])\epsilon = \epsilon$$

as required.

On the boundary of N, or rather, slightly beyond the boundary of N, the first term vanished. We note that the second summand vanishes for $x_1 > 0$ while the third one vanishes for $x_1 < 0$. If $x_1 < 0$ an optimal trajectory could reach the boundary of the disc only in the second quadrant above the switching curve (Figure 4.2). Along such a trajectory $\epsilon = -1$. Once the system exits the circle of radius R, the value of $(R-r)$ becomes negative and the value of ϵ^* becomes $-x_2$. The system would now proceed along an arc of a circle with center at the origin in the counterclockwise direction. As long as the moving point lies above R_{11} the value of ϵ remains -1, and the term

$$(1-\epsilon \operatorname{sgn}[R-r])\epsilon,$$

which appears in the second summand, vanishes. However, once the system, moving as it does on its circular arc, crosses the switching curve, the value of ϵ changes to $+1$. At this juncture the coefficient $\{1 + \epsilon \operatorname{sgn}[R-r]\}$ vanishes whereas $\{1-\epsilon \operatorname{sgn}[R-r]\} = 2$. Thus, the value of ϵ^* is now switched to the value of ϵ, which is $+1$. In other words, once the system reaches the switching curve, control is switched $\epsilon^* = +1$ and the system is steered into the origin along R_{11}.

Similar considerations are obtained in the case when the system reaches the boundary of N in the region $x_1 > 0$, except that in this case the second summand vanishes and it is the third one which furnishes effective control.

(ii) The Case When $1 < R < \sqrt{3}$

Reference is made to Figure 4.3. Consider trajectories starting

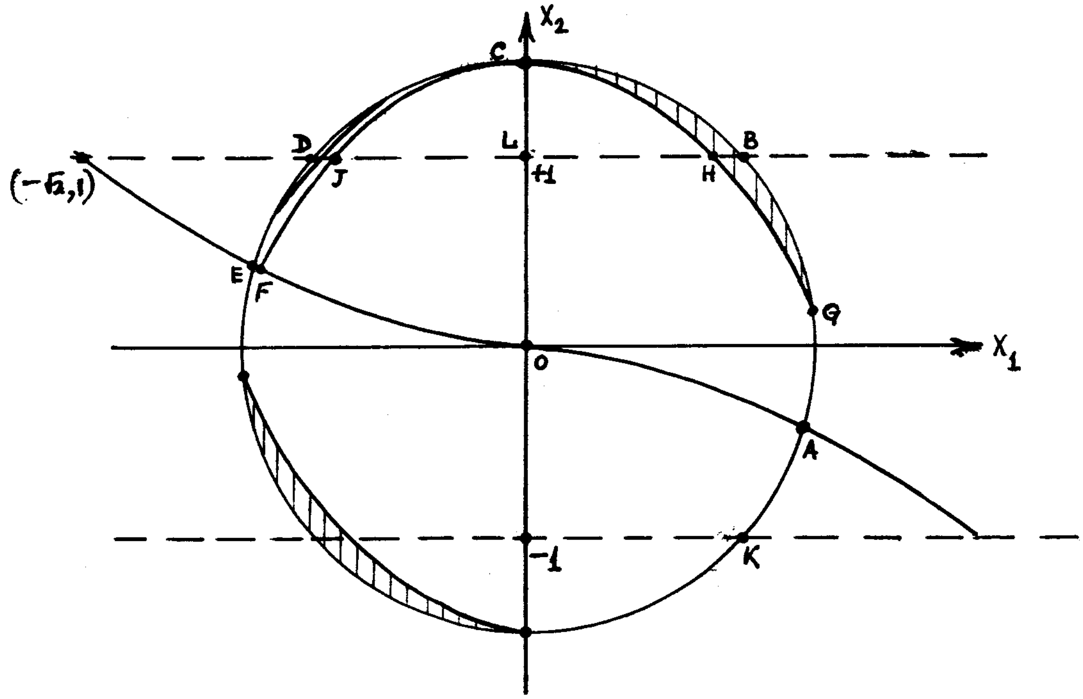


FIGURE 4.3

within N, above the switching curve, in the first or fourth quadrant. Such trajectories (for which $\epsilon = -1$) cannot leave the circle through the arc AB since $\dot{r}$ is decreasing there. Let GHC be the arc of the trajectory, with $\epsilon = -1$, which passes through C. Then any trajectory, with $\epsilon = -1$, which starts in the region CHGBC must meet the boundary of N on the arc BC. Such trajectory, if it were to be controllable within N, would have to proceed along the arc BC, which would require setting $\epsilon = -x_2$. However, $x_2 > 1$ on BC and therefore such control is not admissible. It follows that there exists no optimal control for points in the shaded region which would keep the controlled trajectory within N. The shaded region must therefore lie outside the set N_2.

In fact, it is not difficult to see that the region CHGBC actually lies outside the set N_1. For let $\epsilon = u(t)$ be any admissible control, with $u(t) > -1$. If Γ_1 and Γ_2 are two trajectories emanating from the same initial point P in CHGBC and satisfying, respectively

$$\dot{x}_1 = -1, \quad \dot{x}_2 = x_1$$

and

$$\dot{x}_1 = u(t), \quad \dot{x}_2 = x_1, \quad |u(t)| \leq 1$$

then clearly Γ_2 lies to the right of Γ_1. The trajectory Γ_2 would therefore be forced to the boundary of N somewhere between G and C. Since x_2 is increasing along Γ_2, the system would then have to proceed counterclockwise along an arc of the circle leading to the point C. This, however, is inadmissible. Hence no point in CHGBC is controllable within N even if the condition of optimality is dropped.

Every trajectory emanating in the region OAGHCO, with $\epsilon = -1$, will reach the x_2-axis between O and C and will proceed thence to the third quadrant. In the region CDLC the value of $\dot{r} = (x_1/r)(\epsilon + x_2)$ is negative for $\epsilon = -1$. Hence all trajectories, with $\epsilon = -1$, emanating from or crossing through this region, cannot reach the boundary of the circle along the arc CD. When continued forward in time they may either intersect R_{11} at some point between E and O without ever reaching the boundary of the circle, or they may intersect the circular arc ED before reaching R_{11}. In the first case the value of ϵ is switched to $+1$ and the system proceeds to the origin without ever reaching the boundary of N. In the latter case ϵ is set equal to $-x_2$. Since the arc ED lies below the line $x_2 = 1$, such control is admissible. When the point

E is reached control is switched to + 1. Thus, every point in the region OAGHCDEFO is optimally controllable within N. The situation below the switching curve is completely symmetric (Figure 4.3). Here, then, is a case in which $N \neq N_1$ but $N_1 = N_2$.

(iii) The Case When $R \geq \sqrt{3}$

Reference is made to Figure 4.4. Let ABC be an arc of the trajectory, with $\epsilon = -1$, which passes through C. The reader will easily convince himself that the set N_1 which is controllable within N consists of the unshaded part of the disc. Moreover, $N_1 = N_2$ so that control within N_1 is optimal. The difference between cases (ii) and (iii) is that in the latter case no trajectories emanating from the interior of N_1 can ever reach the boundary of N, whereas in the former case trajectories emanating from the interior of N_1 could reach the boundary along the arcs DE and KA. In case (iii) ϵ need never be set equal to $-x_2$ whereas in case (ii) ϵ must be set equal to $-x_2$ along the aforementioned arcs. In case (iii) we still have $N \neq N_1$, $N_1 = N_2$.

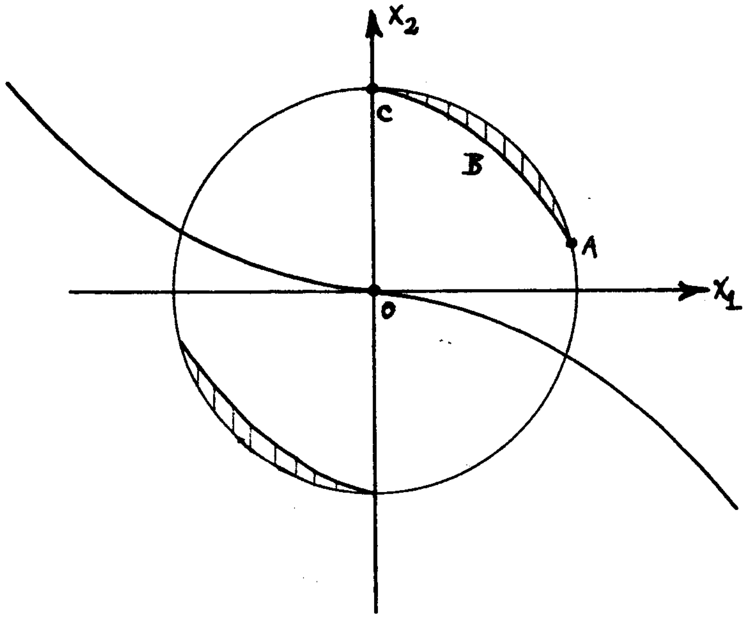


FIGURE 4.4

5. The Subdivisions of ∂N

Let us consider a closed (n-1)-dimensional manifold ∂N topologically equivalent to a sphere. We suppose that ∂N is the boundary of an n-dimensional region N which contains the origin as an interior point. In accordance with a previously explained terminology, we shall say that the region N is controllable within itself, if, for every point $x_o \in N \cup \partial N$, there can be found at least one continuous or piecewise continuous scalar function u(t), whose absolute value does not exceed unity, such that the trajectory of the system

$$dx/dt = Ax + au(t) \tag{5.1}$$

for which $x(0) = x_0$ must pass through the origin for some finite $t = t_0 > 0$ without ever leaving $N \cup \partial N$ for any t between 0 and t_0. Here, as usual, we mean x and a to represent n-vectors, while A is an $n \times n$ constant matrix.

We shall suppose that ∂N is (at least) piecewise representable by equations of the form,

$$f(x) = 0, \tag{5.2}$$

where f is of class C' and is negative for points in N near ∂N and positive for points outside of $N \cup \partial N$ near ∂N. We now contemplate four sets of points S_1, S_2, S_3, S_4 located on ∂N and defined as follows:

S_1 consists of those points on ∂N which "move" outward (that is, away from N) under (5.1) when $u(t) = +1$, and which moves inward under (5.1) when $u(t) = -1$. That is, the points of S_1 are points of egress under $\dot{x} = Ax + a$ and points of ingress under $\dot{x} = Ax-a$, according to a well-known terminology. Analytically, this means that for points in S_1

$$\text{sgn}[(\partial f/\partial x)(Ax \pm a)] = \pm 1 \tag{5.3}$$

S_2 consists of those points on ∂N which move inward under (5.1) when $u(t) = +1$ and which move outward under (5.1) when $u(t) = -1$. That is, the points of S_2 are points of ingress under $\dot{x} = Ax + a$ and points of egress under $\dot{x} = Ax-a$. Analytically this means that for points in S_2

$$\text{sgn}[(\partial f/\partial x)(Ax \pm a)] = \mp 1 \tag{5.4}$$

S_3 consists of those points on ∂N which move inward under (5.1) when $u(t) = \pm 1$. That is, the points of S_3 are points of ingress under both $\dot{x} = Ax + a$ and $\dot{x} = Ax-a$, which means, analytically, that for points in S_3

$$(\partial f/\partial x)(Ax \pm a) < 0 \tag{5.5}$$

S_4 consists of those points on ∂N which move outward under (5.1) when $u(t) = \pm 1$. That is, the points of S_4 are points of egress under both $\dot{x} = Ax + a$ and $\dot{x} = Ax-a$, which means, analytically, that for points in S_4

$$(\partial f/\partial x)(Ax \pm a) > 0 \tag{5.6}$$

Evidently $S_1 \cup S_2 \cup S_3 \cup S_4 \subset \partial N$ and $S_i \cap S_j$ is empty whenever $i \neq j$ $(i,j = 1,2,3,4)$. We shall suppose also that

$\partial N - [S_1 \cup S_2 \cup S_3 \cup S_4]$

may be represented as a finite number of cells of dimensionality $< n-1$.

<u>THEOREM 1.</u> If N is controllable within itself, S_4 is empty.

<u>PROOF:</u> Let $x_o \in S_4$ and suppose it is possible to join x_o with the origin by a trajectory of (5.1) which never leaves $N \cup \partial N$ and for which $|u(t)| \leq 1$. Then, with $x(0) = x_o$, we evidently have

$$(\partial f/\partial x)(Ax_o + au(0)) \leq 0 \tag{5.7}$$

for some $u(0)$ with absolute value less than unity. For evidently (5.7) can not hold for any $u(0) = \pm 1$, because of (5.6). Subtracting the left member of (5.6) (with $x = x_o$) from the left member of (5.7) we also obtain

$$(\partial f/\partial x)a[u(0) \mp 1] < 0 \tag{5.8}$$

Since $\text{sgn}[u(0) \mp 1] = \mp 1$ and since (5.8) holds for <u>both</u> determinations of the ambiguous sign, we find from the upper sign that $(\partial f/\partial x)a > 0$

and from the lower sign that $(\partial f/\partial x)a < 0$, when $x = x_o$. The theorem follows at once from this palpable contradiction.

<u>THEOREM 2.</u> No trajectory of (5.1) initially within N can reach a point of S_3 without first leaving N.

The proof of this theorem is entirely similar to the proof of Theorem 1. It may be formulated as follows:

Suppose there were a point $x_o \in S_3$ through which passes a trajectory of (5.1) at $t = t_o > 0$, i.e., $x(t_o) = x_o$. Now, if $x(t) \in N \cup \partial N$ for all positive $t \leq t_o$, as would have to be the case for some $u(t)$ if the theorem were false, we would have at $x = x_o$

$$(\partial f/\partial x)(Ax_o + au(t_o)) \geq 0 \qquad (5.9)$$

since f is negative within N, zero on ∂N, and positive without N. It is obvious that $|u(t_o)| < 1$, since otherwise (5.5) would be contradicted by (5.9) at $x_o \in S_3$. From (5.5) and (5.9) we also obtain by subtraction

$$(\partial f/\partial x)a[u(t_o) \mp 1] > 0 \qquad (5.10)$$

Since $\operatorname{sgn}[u(t_o) \mp 1] = \mp 1$ and since (5.10) holds for both choices of the ambiguous sign, we find from the upper sign that $(\partial f/\partial x)a < 0$ and from the lower sign that $(\partial f/\partial x)a > 0$ at $x = x_o$. The theorem follows at once from this contradiction.

As a result of Theorems 1 and 2, we can virtually dismiss from further attention the behavior of trajectories on S_3 or S_4.

THEOREM 3. If a trajectory $x(t)$ of (5.1) lies on ∂N throughout a time interval $t_o < t < t_1$, then

$$u(t) = -(\partial f/\partial x)Ax/[(\partial f/\partial x)a] \tag{5.11}$$

Secondly such a piece of some trajectory, with $|u| \leq 1$, passes through every interior point of S_1 or S_2 and, thirdly, $(\partial f/\partial x)a$ can not vanish on S_1 or S_2.

PROOF: Since $x(t)$ lies on ∂N throughout the interval $t_o < t < t_1$, we must have $f[x(t)] \equiv 0$. Differentiating this identity with respect to t and replacing dx/dt by the right hand member of (5.1) we obtain

$$(\partial f/\partial x)[Ax + au(t)] \equiv 0 \tag{5.12}$$

from which we obtain (5.11) immediately, at least, if $(\partial f/\partial x)a \neq 0$. The second assertion of the theorem follows from (5.3) in the case of an interior point x of S_1, or from (5.4) in the case of an interior point of S_2. In either case, the scalar $L(u) = (\partial f/\partial x)[Ax + au]$, considered as a (linear) function of the scalar variable u, changes sign on the interval $-1 \leqq u \leqq +1$. Hence, it must vanish at some intermediate value of u.

The fact that $(\partial f/\partial x)a \neq 0$ anywhere on S_1 or S_2 follows from the fact that the linear function $L(u) \not\equiv$ constant, (otherwise it could not change sign as noted above). Therefore, the coefficient of u in $L(u)$ is not zero. This coefficient is, of course, precisely $(\partial f/\partial x)a$.

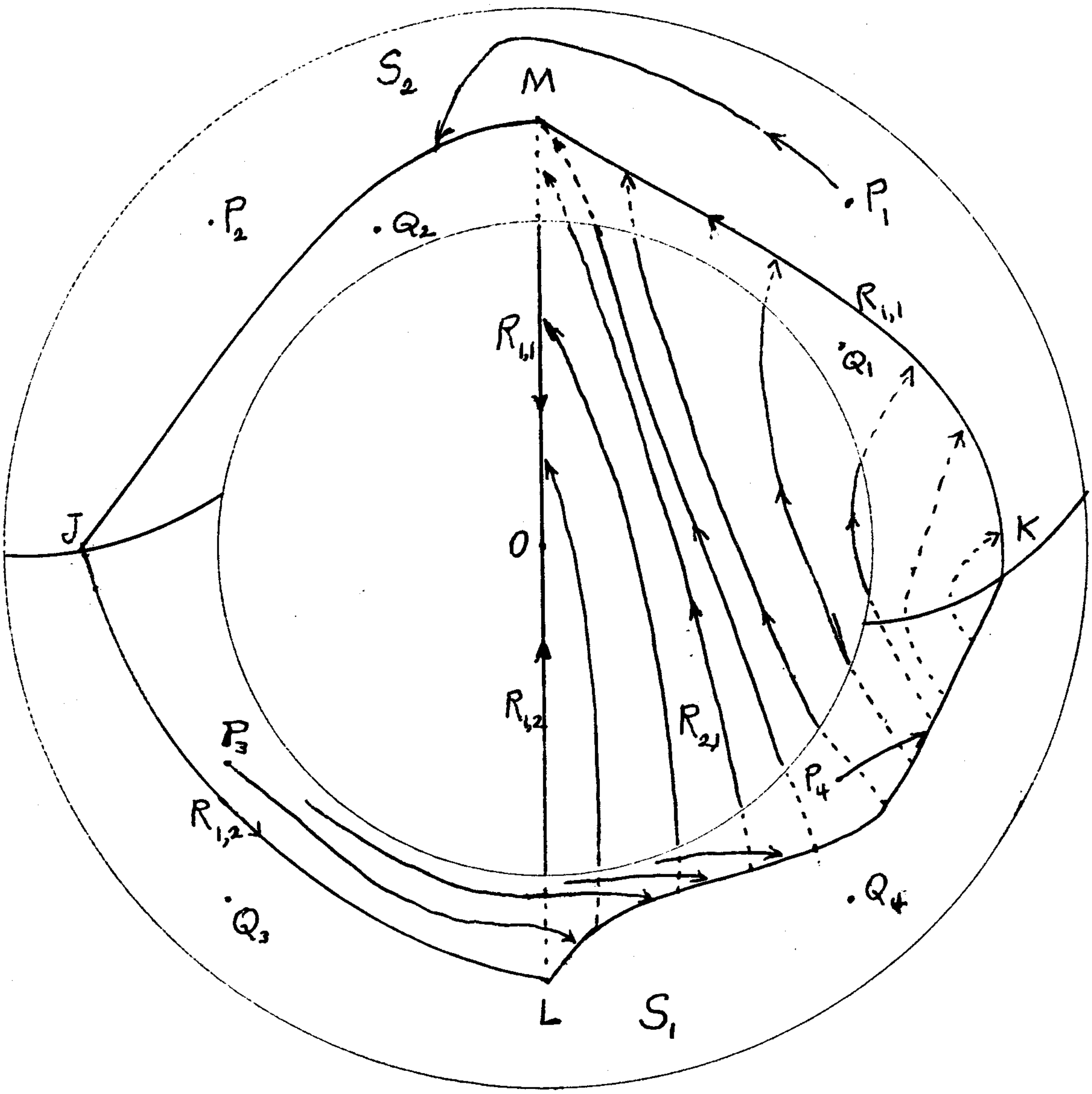


FIGURE 5.1

In this figure (representing a three-dimensional problem) the only part of the boundary shown is the set $S_1 \cup S_2$ represented by the annulus. The leaf $R_{1,1}$ of the switching curve R_1, is represented by OMK, OM being interior to N while $MK \subset S_2$. The leaf

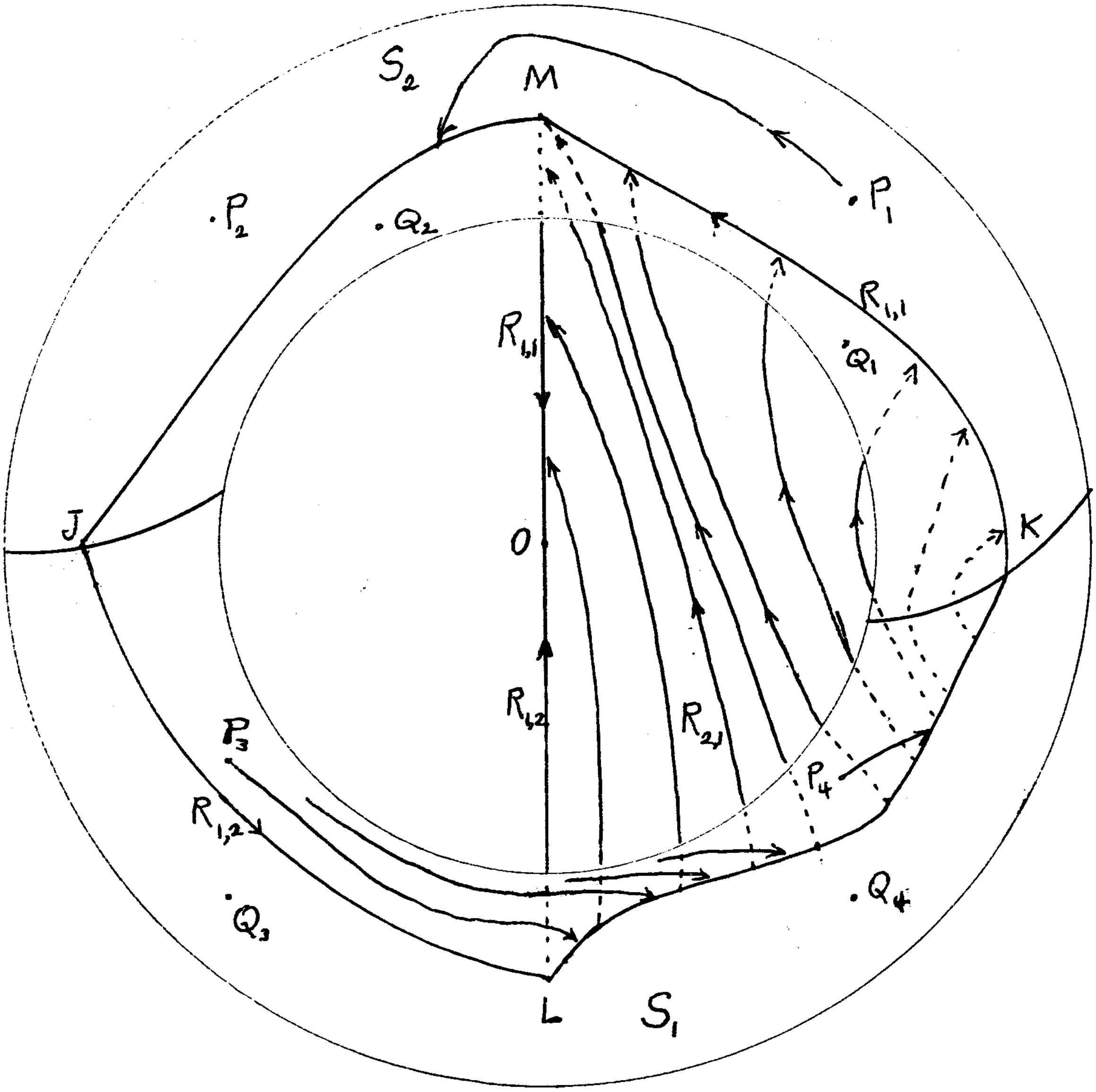


FIGURE 5.1

In this figure (representing a three-dimensional problem) the only part of the boundary shown is the set $S_1 \cup S_2$ represented by the annulus. The leaf $R_{1,1}$ of the switching curve R_1, is represented by OMK, OM being interior to N while $MK \subset S_2$. The leaf

$R_{2,1}$ of the switching surface R_2 consists of those half-trajectories "pointing at" $R_{1,1}$ along its whole length OMK . $R_{2,1}$ intersects S_1 along the curve LK and coincides with S_1 "above" this curve LK. $R_{1,2}$ is represented by OLJ. $R_{2,2}$ is not shown, so as not to clutter up the figure too much. But the boundary of $R_{2,2}$ contains OLJ, and $R_{2,2}$ intersects S_2 along a curve JM and coincides with S_2 above JM. See the text for comments on points $P_1, P_2, Q_1, Q_2 \in S_2$ and $P_3, P_4, Q_3, Q_4 \in S_1$.

6. The Case When $N = N_1 = N_2$

In the sequel we suppose not only that the region N is controllable within itself but that it is strongly controllable within itself in the sense of time optimality. This means that among all the admissible controls yielding a trajectory defined and contained in $N \cup \partial N$ on some interval $0 \leq t \leq t_o$ which connects a given point x_o with the origin $(x(0) = x_o,\ x(t_o) = 0)$, there is always at least one for which t_o is a minimum.

The problem of finding time optimal trajectories is then greatly simplified by a known theorem (referred to in more detail in § 2 of the present chapter) according to which time optimal trajectories must consist (at least, in the present instance, if certain conditions of generality relative to the composition of ∂N are satisfied) of a finite number of arcs of the following three types.

Type 1. Solutions of the system $\dot{x} = Ax + a$ which lie interior to N.

Type 2. Solutions of the system $\dot{x} = Ax-a$ which lie interior to N.

Type 3. Solutions of the system $x = Ax-a[(\partial f/\partial x)Ax][(\partial f/\partial x)a]^{-1}$ which lie in the set $S_1 \cup S_2$.

Any continuous trajectory consisting of a finite number of arcs of these three types which leads from an initial point $x_o \epsilon N \cup \partial N$ to the origin will here be called a bang-bang (constrained) trajectory. The theorem referred to above does not say that bang-bang trajectories are always time optimal but rather that any time optimal trajectory leading from x_o to the origin must be bang-bang. This means that in the search for time optimal trajectories we may limit ourselves to the class of bang-bang trajectories. For this reason the study of bang-bang trajectories is likely to prove fruitful.

Just as in the unconstrained problem, we define switching manifolds of various dimensionalities as the loci of the end points of such arcs of all possible bang-bang trajectories. The parts of switching manifolds involving end points of arcs of Type 3 must lie completely on ∂N and indeed must furthermore lie in $S_1 \cup S_2$ in accordance with Theorem 3 of Section 5. The other parts of the switching manifolds would appear to be somewhat the same as in the unconstrained problem. At least the switching manifolds may theoretically be found by moving backward from the origin just as in the unconstrained problem.

Thus the one-dimensional switching manifold R_1 consists of two leaves $R_{1,1}$ and $R_{1,2}$, the first of which (i.e., $R_{1,1}$) always contains the connected part of the half-trajectory of the system $\dot{x} = Ax + a$ for $t \leqq 0$ which lies within $\bar{N}$ and which is at the origin when $t = 0$; but $R_{1,1}$ also in general contains arcs of Type 3 lying in S_2 (see Figure 5.1, drawn for $n = 3$), the whole of $R_{1,1}$ being a continuous curve joining the origin with a boundary point K of S_2.

Next the leaf $R_{2,1}$ of the two-dimensional switching manifold R_2 always contains the connected parts of all the half trajectories of the equation $\dot{x} = Ax-a$ for $t \leqq 0$ which lie within $\bar{N}$ and which are on $R_{1,1}$ when $t = 0$; but $R_{2,1}$ also in general contains arcs of Type 3 lying on S_1, the whole of $R_{2,1}$ being a continuous surface whose boundary includes $R_{1,1}$ and a curve on the boundary of S_1. Notice that, if $P \in (R_{1,1} \cap \partial N)$, then $P \in S_2$, so that P is a point of egress for the system $\dot{x} = Ax-a$. Thus the half trajectory $t \leqq 0$, which at $t = 0$, is at P, yields points interior to N when $-t$ is small, as required by the above description.

Proceeding by induction, the leaf $R_{k,1}$ of the k-dimensional switching manifold $R_k (1 < k \leq n-1)$ always contains the connected parts of all the half-trajectories of the system $\dot{x} = Ax-(-1)^k a$ for $t \leq 0$ which lie within $\overline{N}$ and which are on $R_{k-1,1}$ when $t = 0$; but $R_{k,1}$ also in general contains arcs of Type 3 lying in $S_{\frac{1}{2}[3-(-1)^k]}$, the whole of $R_{k,1}$ being a continuous k-surface whose boundary includes $R_{k-1,1}$ and a (k-1)-dimensional manifold lying on the boundary of $S_{\frac{1}{2}[3-(-1)^k]}$. Notice that, if $P \in (R_{k-1,1} \cap \partial N)$, then $P \in S_{\frac{1}{2}[3-(-1)^k]}$, so that P is a point of egress for the system $\dot{x} = Ax-(-1)^k a$. Thus the half trajectory $t < 0$ yields points interior to N when $-t$ is small, as required by the above description.

After obtaining the leaf $R_{n-1,1}$ of the switching manifold of highest dimensionality R_{n-1}, the connected parts of all the half trajectories of the system $\dot{x} = Ax-(-1)^n a$ for $t \leq 0$ which lie within $\overline{N}$ and which are on $R_{n-1,1}$ when $t = 0$ make up an n-dimensional region $T_1 \subset \overline{N}$ whose points can be steered along a bang-bang trajectory into the origin via $R_{n-1,1}, R_{n-2,1}, \ldots, R_{1,1}$ using $-(-1)^n$ as the initial value of u, the other values of u being thenceforward uniquely defined. They are, as a matter of fact, either $+1$ or -1 except at certain points on ∂N where they are determined by (5.11) or Section 5.

By repeating the above discussion with the modification that $R_{1,1}$, $R_{2,1}$, $R_{k,1}$, S_1,S_2,a are to be replaced respectively by $R_{1,2}$, $R_{2,2},\ldots$, $R_{k,2}$, $S_2,S_1,-a$ we obtain the other leaves of the switching manifolds, as well as an n-dimensional region $T_2 \subset \bar{N}$ whose points can be steered along a bang-bang trajectory via $R_{n-1,2}$, $R_{n-2,2},\ldots$, $R_{1,2}$ using $(-1)^n$ as the initial value of u, and with the other values of u uniquely determined as before.

There are also certain points, initially on ∂N, which do not begin with $u = \pm 1$, but rather with the value of u given by (5.11) of Section 5. This is because such points are already on a part of a switching manifold which lies on ∂N. Examples of such points are indicated in Figure (5.1) by P_1,P_2,P_3,P_4. On the other hand the initial value to be taken at Q_1 or Q_2 would be $u = +1$, while at Q_3 or Q_4 the initial value of u would be -1.

Evidently there is much lack of rigor in the above discussion. For one thing, although we might conceivably claim that $\bar{N} = \bar{T}_1 \cup \bar{T}_2$ since N is assumed to be strongly controllable, it would probably be more difficult to prove that T_1 and T_2 have no common point. If T_1 and T_2 were to have a non-vacuous intersection, we would have a set of points for which bang-bang control is not unique and this would make it more difficult to decide which control is optimal.

For another thing, it is not entirely clear, for example, why one should terminate $R_{1,1}$ at a boundary point K of S_2 (see Figure 5.1). It is possible that in some problems it might have to be continued into S_1 even though, if this were done, the leaf $R_{2,1}$ would be very peculiar. It would look rather like two leaves joined at the point K and with $u = +1$ instead of -1 on the part near $R_{1,1}$ beyond K.

The situation becomes even more complex when we try to discuss the natural boundaries of the leaves of higher dimensionality.

In attempting to illustrate the above theory, we considered the system $\dot{x}_1 = u(t)$, $\dot{x}_2 = x_1$, $\dot{x}_3 = x_2$, subject to the constraint $x_1^2 + x_2^2 + x_3^2 \leqq r^2$ as well as, of course, $|u| \leqq 1$. That is, we attempted to take $N = (x| \sum_{i=1}^{3} x_i^2 < r^2)$. It was found, however, that such an N is not controllable within itself, no matter how small the positive number r may be chosen; for it was found that the set S_4 is never vacuous if N is chosen in this way. In order to get a set controllable within itself part of the sphere $x_1^2 + x_2^2 + x_3^2 \leqq r^2$ must be discarded. Such an example is discussed in the following chapter.

CHAPTER 20

A THREE DIMENSIONAL EXAMPLE OF BANG-BANG CONTROL WITH PHASE COORDINATE CONSTRAINTS

A Three Dimensional Example Of Bang-Bang Control With Constraints

We consider the system $\dot{x}_1 = u$, $\dot{x}_2 = x_1$, $\dot{x}_3 = x_2$, subject to $|u| \leqq 1$ along with the two further constraints $3x_1 + x_2^2 + x_3^2 - 1 \leqq 0$ and $-3x_1 + x_2^2 + x_3^2 - 1 \leqq 0$. It will be somewhat more convenient, however, to use x, y, z in place of x_2, x_3, x_1, respectively. The system is therefore written in the form $\dot{z} = u$, $\dot{x} = z$, $\dot{y} = x$ and the constraint conditions are $|u| \leqq 1$, $3z + x^2 + y^2 - 1 \leqq 0$ and $-3z + x^2 + y^2 - 1 \leqq 0$. The last two conditions mean that the motion is required to take place within, or on the boundary of, the three-dimensional region N bounded by the two paraboloids of revolution $z = (\frac{1}{3})(1-x^2 - y^2)$ and $z = -(\frac{1}{3})(1-x^2 - y^2)$. The part of the boundary, for which $z > 0$, lies on the first of these paraboloids and will be referred to as the "upper cap." The part of the boundary, for which $z < 0$, lies on the second paraboloid, and will be referred to as the "lower cap." The only other boundary points of the solid are the points of the unit circle $x^2 + y^2 = 1$ in the plane $z = 0$.

The upper cap is, in this example, identical with the set S_1, defined in Section 5 of Chapter 19 while the lower cap is the set S_2 (also defined there). That is, every point of the upper cap is a point of egress for the system $\dot{z} = +1$, $\dot{x} = z$, $\dot{y} = x$, and is a point of ingress for the system $\dot{z} = -1$, $\dot{x} = z$, $\dot{y} = x$. The reverse is true of the lower cap.

To prove these assertions we merely note that

$$\frac{d}{dt}[3z + x^2 + y^2 - 1] = 3u + 2xz + 2yx \tag{1}$$

which on the upper cap may also be written in the form,

$$\frac{d}{dt}[3z + x^2 + y^2 - 1] = 3u + (\tfrac{2}{3})x(1-x^2-y^2) + 2xy \tag{2}$$

Since the maximum absolute value on the cap of each of the quantities $(1-x^2-y^2)$, x and y, is 1, it is seen at once that the sign of $\frac{d}{dt}[3z + x^2 + y^2 - 1]$ when $|u| = 1$ is the same as the sign of u. Similar considerations apply to the lower cap.

More generally one might consider the paraboloids $\pm cz + x^2 + y^2 = a^2$. We took $a = 1$, $c = 3$, in order to have a simple example of a case in which the upper cap is S_1 and the lower cap is S_2. This would not be the case, when c is sufficiently small compared with a. For instance, when $a = 1$ and $c = 1$, the point $x = -\sqrt{\frac{5}{12}}$, $y = \frac{1}{2}$, $z = \frac{1}{3}$, is on the upper cap and yet is a point of ingress for the system $\dot{z} = +1$, $\dot{x} = z$, $\dot{y} = x$, instead of being a point of egress as it should be if the entire upper cap contained only points of S_1.

Returning to the case $c = 3$, $a = 1$, (to which we hereafter confine attention), it may be mentioned immediately that the region N, as specified, is not controllable within itself. In fact, the points $(0, \pm 1, 0)$ can not be moved by taking $|u| \leq 1$, without leaving the region. For we find from (1) that $\frac{d}{dt}[3z + x^2 + y^2 - 1] = 3u$, if $x = 0$, $y = \pm 1$, and $z = 0$. Hence, if we are to remain on or below the paraboloid of the upper cap we must take $u \leqq 0$. On the other hand we have $\frac{d}{dt}[-3z + x^2 + y^2 - 1] = -3u$, if $x = 0$, $y = \pm 1$, $z = 0$, so that, if we are to remain on or above the paraboloid of the lower cap, we must take $u \geqq 0$. Hence the only possible way to stay within N or on ∂N is to take $u = 0$. But for this value of u the point in question is an equilibrium point. It will appear in the sequel that many other points of N and its boundary must be discarded if we are to be left with a region controllable within itself. We shall describe how such a discard may be made so that the remaining region will be controllable within itself using only bang-bang trajectories in the generalized sense defined in Section 6 of the preceding chapter. The arcs of these bang-bang trajectories are of three types, namely:

TYPE 1. Solutions of $\dot{z} = +1$, $\dot{x} = z$, $\dot{y} = z$.

TYPE 2. Solutions of $\dot{z} = -1$, $\dot{x} = z$, $\dot{y} = x$.

TYPE 3a. Solutions of $\dot{z} = -(\frac{2}{3})x(y + z)$, $\dot{x} = z$, $\dot{y} = x$.

TYPE 3b. Solutions of $\dot{z} = +(\frac{2}{3})x(y+z)$, $\dot{x} = z$, $\dot{y} = x$.

Arcs of Types 1 and 2 are allowed only interior to N although their end points may lie on the boundary of N. Arcs of Type 3a occur only on the upper cap, while arcs of Type 3b occur only on the lower cap. Bang-bang trajectories are made up of a finite number of arcs of these three types exclusively.

Arcs of Type 3a may be adequately discussed by considering their orthogonal projections on the plane $z = 0$. These latter curves satisfy the differential equations

$$\begin{aligned}\dot{x} &= (\tfrac{1}{3})(1-x^2-y^2)\\ \dot{y} &= x \qquad (3)\end{aligned}$$

Similarly the arcs of Type 3b are projected onto the plane $z = 0$ and these projected curves satisfy the differential equations,

$$\begin{aligned}\dot{x} &= -(\tfrac{1}{3})(1-x^2-y^2)\\ \dot{y} &= x \qquad (4)\end{aligned}$$

It is easy, by direct inspection of the differential equations, to discuss the phase portraits of (3) and (4). Thus the trajectories of (3) cross the y-axis with zero slopes; they cross the unit circle with infinite slopes (except at the singular points $x = 0$, $y = \pm 1$); within the unit circle they have positive slopes to the right of the y-axis, and they have negative slopes to the left of the y-axis. The situation is reversed outside the unit circle, but it is only the interior of the unit circle with which we are primarily concerned. If, for the moment, we do consider points outside as well as inside the unit circle, we may note the fact that the singular point $(0,-1)$ is a saddle point and the point $(0, +1)$ is a center. This leads to a situation illustrated in Figure 1. Similarly the trajectories on the lower cap are projected onto the curves on the plane $z = 0$ illustrated in Figure 2.

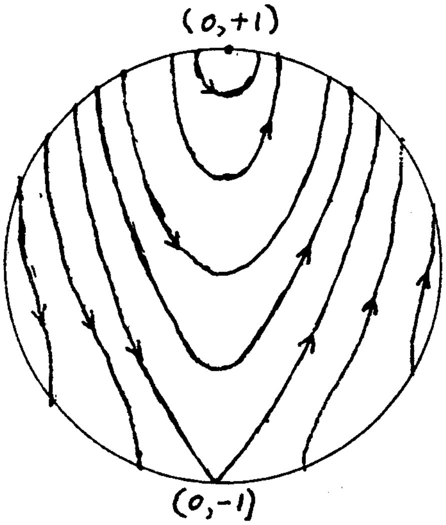


FIGURE 1. (upper cap)

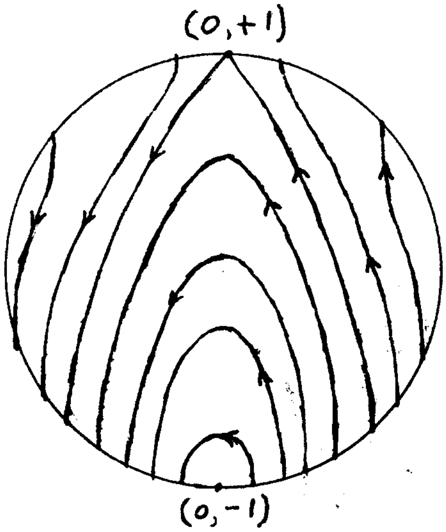


FIGURE 2. (lower cap)

Actually systems (3) and (4) admit simple exponential integrating factors, namely $e^{(2/3)y}$ and $e^{-(2/3)y}$, respectively, so that we readily find explicit equations for the curves in Figure 1. They are of the form

$$e^{(2/3)y}[x^2 + y^2 - 3y + \tfrac{7}{2}] = \text{const.} \qquad (5)$$

The corresponding equations for the curves in Figure 2 have the form,

$$e^{-(2/3)y}[x^2 + y^2 + 3y + \tfrac{7}{2}] = \text{const.} \qquad (6)$$

Not all of the curves on Figures 1 and 2 have equal importance. In our discussion of the more important of these curves, we shall

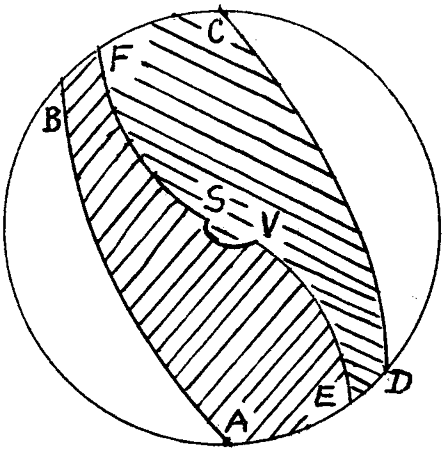


FIGURE 3. (upper cap)

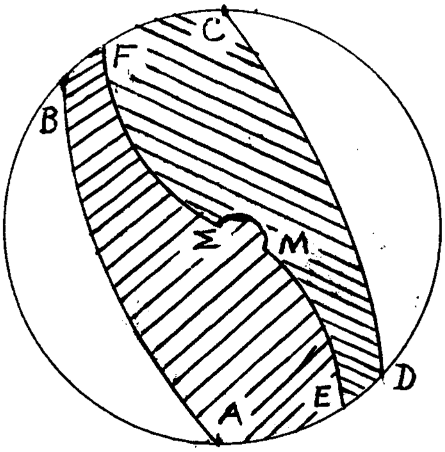


FIGURE 4. (lower cap)

need to clutter up the figures with other markings. We therefore refer to Figures 3 and 4, where only the two arcs BA and FS in Figure 3 satisfy the differential equations (3) or the integrated equation (5), and where only the two arcs DC and EM in Figure 4 satisfy the differential equations (4) or the integrated equations (6).

The point S in Figure 3 is determined in such a manner that it is the projection on the plane $z = 0$ of the point P where the trajectory of Type 2 passing through the origin intersects the upper cap. The coordinates of P may be found by solving a certain algebraic equation and in fact are found to be approximately $x = -.0552$, $y = +.0061$, $z = +.3323$. Thus the point S on the xy-plane is $(-.0552, y = +.0061)$. Using these values, the constant on the right hand side of equation (5), was found to be approximately 3.4990 for the curve FS. After this, we may solve a simple transcendental equation to find the coordinates of $F(-.5597, +.8287)$.

Similarly the point M in Figure 4 is the projection on the plane $z = 0$ of the point where the trajectory of Type 1 passing through the origin intersects the lower cap. It turns out that M has the coordinates $+.0552$ and $-.0061$, and the curve EM cuts the unit circle at the point $E(+.5597 , -.8287)$.

The curve BA in Figure 3 is the trajectory of Type 3a which approaches the saddle point $A(0,-1)$ of the system (3). The constant on the right hand side of equation (5) corresponding to BA is therefore $(\frac{15}{2})e^{-2/3}$ and it is then possible to find the coordinates of B $(-.7231 + .6906)$ by solving an appropriate transcendental equation.

The curve DC in Figure 4 is the trajectory of Type 3b which approaches the saddle point $C(0, +1)$ of the system (4). Proceeding as before it is possible to find the coordinates of $D(+.7231, -.6906)$.

From this description and from the material in Section 6 of the preceding chapter, it is evident that the leaf $R_{1,2}$ of the one-dimensional switching manifold consists of the curve on the upper cap corresponding to the curve FS of Figure 3 along with an arc of Type 2 connecting with the origin. The leaf $R_{1,1}$ consists of the curve on the lower cap corresponding to the curve EM of Figure 4 along with an arc of Type 1 connecting with the origin.

The leaf $R_{2,1}$ of the two-dimensional switching manifold consists of all trajectories leading into $R_{1,1}$ of a certain kind. These trajectories, just before their junctions with $R_{1,1}$, will run along arcs of Type 2; but, if these arcs are followed backward, it will be found that they intersect the upper cap along a certain curve whose projection onto the plane $z = 0$ is the curve of Figure 3, EVS. Part of EVS, namely VS, represents the intersections of the upper cap with those arcs of Type 2 which join onto $R_{1,1}$ at interior points of N. The other part of EVS, namely EV, represents the intersection of the upper cap with those arcs of Type 2 which join onto $R_{1,1}$ at boundary points of N. As mentioned above the part of $R_{1,1}$ on the boundary of N is represented by the curve EM in Figure 4. The curve EV in Figure 3 thus represents a curve on the upper cap whose points are carried along arcs of Type 2 to points on a curve on the lower cap represented by the curve EM in Figure 4. But $R_{2,1}$ consists not only of the arcs of Type 2, just mentioned, but also the arcs of Type 3a which fill out an area on the upper cap, whose projection, on the plane $z = 0$, is the shaded region AEVSFBA in Figure 3.

The leaf $R_{2,2}$ of the two-dimensional switching manifold consists of all trajectories leading into $R_{1,2}$ of a certain kind. These trajectories, just before their junctions with $R_{1,2}$ will run along arcs of Type 1; but, if these arcs are followed backward, it will be found that they intersect the lower cap along a certain curve whose projection onto the plane $z = 0$ is the curve $F \Sigma M$

of Figure 4. Part of $F \Sigma M$, namely ΣM, represents the intersection of the lower cap with those arcs of Type 1 which join onto $R_{1,2}$ at interior points of N. The other part of $F \Sigma M$, namely $F\Sigma$, represents the intersection of the lower cap with those arcs of Type 1 which join onto $R_{1,2}$ at boundary points of N. As previously indicated, the part of $R_{1,2}$ on the boundary of N is represented by the curve FS in Figure 3. The curve $F\Sigma$ in Figure 4 thus represents a curve on the lower cap whose points are carried along arcs of Type 1 to points on a curve on the upper cap represented by the curve FS in Figure 3. But $R_{2,2}$ consists not only of the arcs of Type 1, just mentioned, but also the arcs of Type 3b which fill out an area on the lower cap whose projection, on the plane $z = 0$, is the shaded region $EDCF\Sigma ME$ in Figure 4.

To complete our description of Figures 3 and 4, it remains to define the curve DC in Figure 3 as the orthogonal projection on $z = 0$ of the curve of intersection of the upper cap with the arcs of Type 2 leading into the points on the lower cap represented by the curve DC in Figure 4. Similarly the curve BA in Figure 4 is the orthogonal projection on $z = 0$ of the curve of intersection of the lower cap with the arcs of Type 1 leading into the points on the upper cap represented by the curve BA in Figure 3. Moreover, the points on the upper cap represented by shading with lines of negative slope in Figure 3 are carried along arcs of Type 2 into the points of the lower cap represented by shading with lines of negative slope in Figure 4, except for some which reach points of $R_{2,2}$ interior to N before reaching the boundary of N. Similarly, the points on the lower cap represented by shading with lines of positive slope in Figure 4 are carried along arcs of Type 1 into the points of the upper cap represented by shading with lines of positive slope in Figure 3, except for some which reach points of $R_{2,1}$ before reaching the boundary of N.

We are now in a position to isolate a subset N_1 of N which is controllable within itself by bang-bang control. Namely N_1 is bounded above by the part of the upper cap whose projection on the xy-plane is the shaded region in Figure 3; it is bounded below by the part of the lower cap represented by the shaded region in Figure 4; it is bounded laterally on the left by arcs of Type 1 leading from the curve BA on the lower cap (whose orthogonal projection is represented in Figure 4) to the curve BA on the upper cap (whose orthogonal projection is represented in Figure 3); and, finally, it is bounded laterally on the right by arcs of Type 2 leading from the curve D C on the upper cap to the curve DC on the lower cap. We must exclude from N_1 the points on these lateral boundaries as well as the cruves BA and DC on both the upper and lower caps. For our bang-bang control sends all such points asymptotically into one or the other of the singular points A and C. Other points in $N-N_1$ are probably completely uncontrollable, although this has not been proved.

We now give a preliminary description of how bang-bang control is effected within N_1.

If the point is initially on the part of the upper cap represented by shading with lines of negative slope in Figure 3 or if it is initially slightly below this region, the point is carried along an arc of Type 2 until it meets $R_{2,2}$. If the meeting with $R_{2,2}$ occurs on the boundary, that is, on the part of the lower cap represented by similar shading in Figure 4, the system is switched to an arc of Type 3b until it meets the curve represented by $F \Sigma M$, at which instant it is switched to an arc of Type 1. If, however, the first meeting with $R_{2,2}$ occurs at an interior point of N_1 the switch to an arc of Type 1 is effected immediately. In either case the point is conveyed along this arc of Type 1 until it meets $R_{1,2}$. This meeting may occur either on the upper cap on the curve represented in Figure 3 by FS or at an interior point of $R_{1,2}$. In the former case the switch to the arc FS of Type 3a is made and then a last switch to an arc of Type 2 is made. In the latter case the switch to the arc of Type 2 is made immediately.

If the point is initially on the part of the upper cap represented by shading with lines of positive slope in Figure 3, the point is carried along an arc of Type 3a, until it meets the curve represented by EVS in Figure 3, at which instant it switches to an arc of Type 2 until it meets $R_{1,1}$, and then is carried into the origin in an obvious way.

If the point is initially just below this region, it must of course, be carried first along an arc of Type 1 until it reaches this region, and then its subsequent motion is the same as that discussed in the preceding paragraph.

The above discussion applies to all points starting on the upper cap in N_1 or just below the upper cap. The discussion for points starting on or just above the lower cap is carried out in an analogous way and is left to the reader.

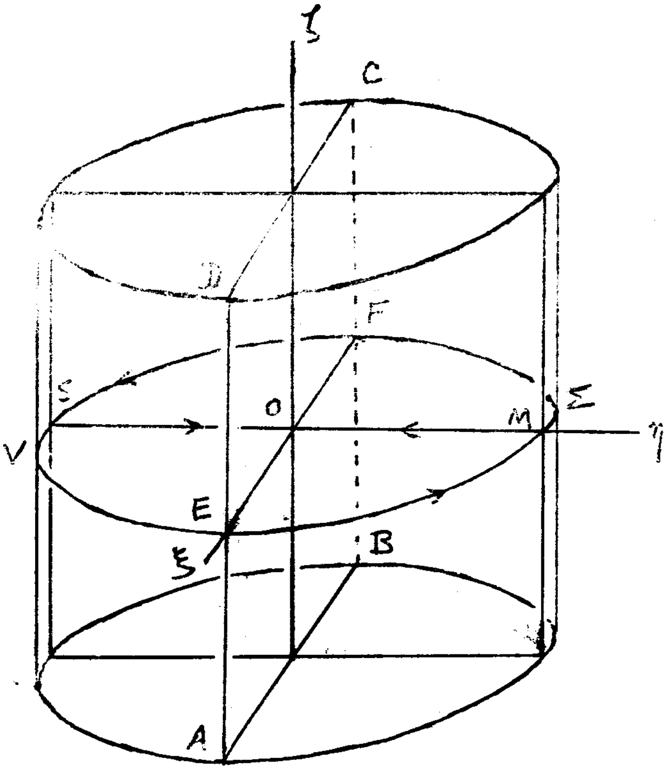


FIGURE 5

For points starting deep in the interior of N_1, one must of course, always start with an arc of Type 1 or 2 depending on which "side" of the two-dimensional switching manifold the initial point may happen to be on. This will be made more clear in the sequel.

It may be seen that it is possible and useful to apply a topological transformation to N_1 in such a manner that it appears as the right circular cylinder $\xi^2 + \eta^2 \leq 1$, $|\zeta| < 1$ in ξ, η, ζ-space.

See Figure 5, which has been drawn and lettered in such a manner that all lettered points in Figure 5 are the topological images of the points on the upper or lower caps whose projections on the plane $z = 0$ are similarly letered in Figures 3 and 4. The topological transformation is further chosen so that the upper cap now appears as the left side of the cylindrical surface, and the laterial boundaries previously referred to now appear as the two bases of the cylinder. The leaves of the switching manifolds are represented in Figure 5 as follows:

$R_{1,1}$: EMO, i.e., points (ξ,η,ζ) such that $\xi^2+\eta^2=1$, $\zeta = 0$, $\xi \geqq 0$, $\eta > 0$, or such that $\xi = 0$, $\zeta = 0$, and $0 < \eta \leqq 1$.

$R_{1,2}$: FSO, i.e, points (ξ,η,ζ) such that $\xi^2+\eta^2=1$, $\zeta = 0$, $\xi \leqq 0$, $\eta < 0$ or such that $\xi = 0$, $\zeta = 0$ and $0 > \eta \geqq -1$.

$R_{2,1}$: All points (ξ,η,ζ) for which either $\zeta = 0$, $\xi^2+\eta^2 < 1$, $\xi > 0$ or for which $\eta < 0$, $\zeta < 0$, and $\xi^2+\eta^2 = 1$.

$R_{2,2}$: All points (ξ,η,ζ) for which either $\zeta = 0$, $\xi^2+\eta^2 < 1$, $\xi < 0$ or for which $\eta > 0$, $\zeta > 0$, and $\xi^2+\eta^2 = 1$..

If the topological transformation which satisfies the above requirements could be given explicitly, it would not be difficult to give an explicit representation of the switching function. For there are essentially six possibilities for the position of the initial point and with each of these possibilities there is a unique choice for the Type of initial arc. These various possibilities and corresponding types are indicated as follows:

1. Interior to cylinder and above $\xi\eta$-plane. Type 2.
2. Interior to cylinder and below $\xi\eta$-plane. Type 1.
3. On upper right cylindrical surface. Type 3b.
4. On lower right cylindrical surface. Type 1.
5. On upper left cylindrical surface. Type 2.
6. On lower left cylindrical surface. Type 3a.

Finally, it may be mentioned that the shaded part of the boundary of N_1, may also be represented as an annulus as in Figure 5.1 of Chapter 19. However, the behavior along the curves where the two caps have common boundaries (i.e., the two halves of the annulus) is somewhat different from that indicated in that figure of Chapter 19.

APPENDIX TO PART A

PARAMETRIC EQUATIONS OF THE THREE-DIMENSIONAL SWITCHING MANIFOLD FOR THE FOURTH ORDER LINEAR SYSTEM WITH EIGENVALUES $(0, 0, -\lambda, \lambda)$

By

Peter S. Ying

Parametic Equations Of The Three-Dimensional Switching Manifold For The Fourth Order Linear System With Eigenvalues $(0, 0, -\lambda, \lambda)$

As we mentioned in Chapter 17, the computation of the three-dimensional switching manifold $R_{3,1}$ of the fourth order linear system with eigenvalues $(0,0,-\lambda,\lambda)$ involves the use of the transformation

$$\begin{aligned}
z_1 &= -y_1 \\
z_2 &= -1-e^{\lambda y_1}[(y_2+1)e^{\lambda y_1} -2] \\
z_3 &= +1-e^{-\lambda y_1}[(y_3-1)e^{-\lambda y_1} +2] \\
z_4 &= -(y_4+y_1^2)
\end{aligned} \qquad (1)$$

We substitute into

$$R_{2,1}\begin{cases}
z_2z_3+z_2^2z_3+z_2^2 < 0 \\
e^{\lambda z_1} < (z_2-z_3-z_2z_3)/2z_2 \\
(z_3-z_2+z_2z_3)^2+4z_2z_3 = 0 \\
z_4+\frac{1}{\lambda^2}\log^2(\frac{z_2-z_3-z_2z_3}{2z_2}) = 0
\end{cases} \qquad (2)$$

and let

$$u = e^{\lambda y_1} \qquad \xi = y_2 + 1$$
$$\eta = y_3 - 1 \qquad \zeta = \lambda^2 y_4 \tag{3}$$

Then (2) becomes

$$-2\xi^2 u^5 + (8\xi + 2\xi^2)u^4 + (\xi^2\eta - 11\xi - 8)u^3 + (-4\xi\eta + 2\xi + 14)u^2$$
$$+ (\xi\eta + 4\eta - 5)u - 2\eta > 0 \tag{4a}$$

$$\frac{2\xi u^2 + (\xi\eta - 3)u - 2\eta}{2(\xi u^2 - 2u + 1)} > 1 \tag{4b}$$

$$(4\xi^2 - 4\xi)u^4 + (4\xi^2\eta - 4\xi + 8)u^3 + (\xi^2\eta^2 - 10\xi\eta - 11)u^2$$
$$+ (4\eta - 4\xi\eta^2 + 8)u + (4\eta^2 + 4\eta) = 0 \tag{4c}$$

$$\zeta = \log^2 \frac{2\xi u^2 + (\xi\eta-3)u-2\eta}{2(\xi u^3-2u^2 + u)} - \log^2 u \qquad (4d)$$

The previous method for the computation of $R_{3,1}$ was to eliminate u between equations (4c) and (4d). On account of the extreme difficulty of this elimination, it seemed wise to investigate the possibility of bypassing this elimination with the purpose of developing an adequate description of this switching manifold, leading perhaps to satisfactory approximations.

Equation (4c) is a quartic equation in u, but, from the definition of u in equations (3) only positive roots are to be accepted. Also, since the real logarithmic function is defined only for positive values of the independent variables, we see that ξ, η, and u also must satisfy

$$1 + \frac{(\xi\eta + 1)u-2(\eta + 1)}{2(\xi u^2-2u + 1)} > 0 \qquad (5)$$

Obviously, if we can solve equation (4c) for u and substitute into equation (4d), we shall obtain a relation between ξ, η and ζ,

[redacted]

which will define a surface in (ξ,η,ζ) - space. This surface does not intersect the plane $\eta = 0$ at points where $\xi < 0$ or where $\xi > 1$.

To prove the underlined statement we set $\eta = 0$ in (4c), thus obtaining

$$f(u) = (4\xi^2 - 4\xi)u^3 + (-4\xi + 8)u^2 - 11u + 8 = 0 \qquad (6)$$

or $u = 0$. This last possibility is excluded by (3). Since the discriminant of equation (6) is

$$\Delta = \frac{1}{16\xi^4(\xi-1)^4}\,[1728\xi^4 - 5168\xi^3 + 5796\xi^2 - 2889\xi + 540],$$

it can be shown that $\Delta > 0$ except when ξ lies on the closed interval between the two real zeros of Δ which are approximately $+ 0.740741$ and $+ 0.749984$. We recall that when $\Delta < 0$, equation (6) has three real roots; while, if $\Delta > 0$, it has only one real root. Suppose now that $\xi < 0$ or $\xi > 1$. Then the coefficient of u^3 in (6) must be positive. Hence $f(-\infty) = -\infty$ while $f(0) = 8 > 0$. Hence (6) has a negative real root, and this is the only real root since $\Delta > 0$ for the values of ξ under consideration. Hence, if $\eta = 0$ and $\xi < 0$ or $\xi > 1$, the equation (4c) has no positive real root. This completes the proof of the underlined statement.

When $\eta = 0$ and $0 < \xi < 1$, the coefficient of u^3 is negative. Hence $f(+\infty) = -\infty$ and $f(0) = 8 > 0$. Hence equation (6) has at least one positive root if ξ is between 0 and 1. It, of course, has three positive roots if

$$.740741 < \xi < .749984.$$

We carried out a machine calculation of the roots of equation (4c) and the values of ζ given by equation (4d) for all even integral values of ξ and η from -20 to + 20 and more detailed calculations in the region $-2 \leqq \xi \leqq +2$ and $-2 \leqq \eta \leqq +2$. From these calculations it was conjectured that there are no points on the surface with $\xi < 0$ for arbitrary values of η.

PART B

SIMULATION AND COMPUTATION

by

J. Schlesinger

J. Gilchrist

K. Ivey

G. Campbell

TABLE OF CONTENTS

PART B

SIMULATION AND COMPUTATION

I. Simulation Of The Third-Order System

(a) The system described by

$$\dddot{x} = \epsilon, \; \epsilon = \pm 1,$$

can be represented in phase space by the coordinates $(\ddot{x}, \dot{x}, x)$. The system moves according to the equations (derived by integration)

$$\ddot{x} = \epsilon t + \ddot{x}(0)$$

$$\dot{x} = \frac{1}{2} \epsilon t^2 + \ddot{x}(0)t + \dot{x}(0)$$

$$x = \frac{1}{6} \epsilon t^3 + \frac{1}{2} \ddot{x}(0)t^2 + \dot{x}(0)t + x(0)$$

Given a point in phase space, the problem is to find the time-optimal path to the origin $\ddot{x} = \dot{x} = x = 0$. This should be accomplished by moving a certain time t_1, then switching ϵ to $-\epsilon$, moving a second time t_2, then switching back to ϵ again, and moving a third time t_3 to 0. The question is to determine the switching times $\{t_1, t_2, t_3\}$. This can be done using the control function (see equation (12A) in Chapter 13)

$$\sigma = -(\text{sgn}\ \sigma_3)\sigma_3^2 - \sigma_2^3, \quad \text{where}$$

$$\sigma_2 = \dot{x} + \frac{1}{2} \ddot{x}^2 \ \text{sgn}[\dot{x} + \frac{1}{2} \ddot{x}^2 \ \text{sgn}\ \ddot{x}]$$

$$\sigma_3 = x + \dot{x}\ \ddot{x}\ \text{sgn}[\dot{x} + \frac{1}{2} \ddot{x}^2 \ \text{sgn}\ \ddot{x}] + \frac{1}{3} \ddot{x}^3$$

By setting ϵ initially to $\text{sgn}(\sigma)$, the plant moves along until it intersects the first switching surface, at which point $\sigma = 0$. ϵ is then switched. The system now moves along a second path until σ is again 0, when ϵ is again switched. The point now moves into the origin monotonically, and one needs merely note the time at which it actually passes through 0.

The control law was simulated on an IBM 1620 computer as follows:

A point $(\ddot{x}, \dot{x}, x)$ is given, and ϵ is computed for the initial value. Then t is incremented from 0 until σ changes sign. It is assumed that the point has just passed through the switching surface. We record t, then set it back to 0, switch ϵ, and proceed along the new path until σ again switches sign. Recording the second t, switching ϵ, setting t back to 0, we follow the last curve until we observe that $|(\ddot{x}, \dot{x}, x)|$ has begun to increase, at which point we assume the point is at its closest to the origin.

In order to test the program, we had to determine check points for which the optimal paths were known. A program to accomplish this involved starting at the origin and marching backwards, switching, marching backward again, switching again, and finally backtracking until we arrived at a point lying (arbitrarily) on the unit sphere. Since the system is symmetric through the origin, we computed only points whose initial ϵ was $+1$, for their negatives have the same times with starting $\epsilon = -1$. We computed some two hundred points, all lying in the same "hemisphere."

The accuracy of the control law simulation depended, of course, on the size of the t-increments.

Because of the "steepness" of the first switching surface, a relatively small overshoot of t_1 resulted in missing the second and third by a rather large amount. When the first t is good, the others are good also.

Δt	POINT	FINAL POINT	T COMPUTED	T ACTUAL
.1	-.93, .43, -.14	-.03, -.03, -.00	1.1, .4, .2	.95, .1, .07
.1	-. 8, .52, -.32	.0, -.05, .00	.8, .5, .5	.75, .4, .45
.1	.02, -.88, .48	.02, -.23, .03	1.1, 1.9, .8	1.05, 1.6, .53
.05	-.95, .41, -. 1	.0, -.02, .0	1.2, .45, .20	1.2, .40,.15
.05	.625, -. 7, .41	.03, .08, .03	.20, .45, .85	.20, .40,.83
.02	-.25, .29, -.94	.01, -.11, .00	.82, 1.4, .84	.80, 1.3, .75
.01	-.93, .43, -.14	.0, .0, .0	1.1, .34, .17	.95, .10,.07
.01	.625, -. 7, .41	.01, .02, .01	.19, .41, .84	.20, .40,.83
.01	.02, -.88, .48	.0, -.03, .0	1.06, 1.64, .56	1.05, 1.6, .53

(b) The third order system

$$\dddot{x} + \lambda^2\dot{x} = \epsilon, \quad \epsilon = \pm 1$$

whose eigenvalue $\lambda = 0, \pm 1,$ is a generalization of the system

$$\dddot{x} = \epsilon,$$

which was described in Section I(a). We concerned ourselves with the new case $\lambda = \pm 1$. After transforming to the space (x_1, x_2, x_3), the system is

$$\dot{x}_1 = \epsilon$$
$$\dot{x}_2 = x_2 + \epsilon$$
$$\dot{x}_3 = -x_3 + \epsilon,$$

and the solution is

$$x_1 = \epsilon t + C_1$$
$$x_2 = C_2 e^t - \epsilon$$
$$x_3 = \epsilon - C_3 e^{-t} .$$

By means of the control function σ:

$$\epsilon = \mathrm{sgn}(\sigma),$$

where $\sigma = -(1 + \sigma_1)F(\sigma_2,\sigma_3) + (1-\sigma_1)G(\sigma_2,\sigma_3)$

$$F(\xi_2,\xi_3) = \{[\xi_3(1 + \xi_2)-\xi_2]^2 + 4\xi_2\xi_3\}\cdot\{\xi_3(1 + \xi_2)^2-\xi_2^2 + \xi_2+ 2(-\xi_2)^{\frac{3}{2}}\}$$

$$G(\xi_2,\xi_3) = \{[-\xi_3(1 - \xi_2) + \xi_2]^2 + 4\xi_2\xi_3\}\cdot\{-\xi_3(1-\xi_2)^2-\xi_2^2 -\xi_2+ 2\xi_2^{\frac{3}{2}}\}$$

$$\sigma_2 = h_2(x_1,x_2,x_3,\sigma_1)$$

$$\sigma_3 = h_3(x_1,x_2,x_3,\sigma_1)$$

$$\sigma_1 = -\mathrm{sgn}\, h_2(x_1,x_2,\epsilon_1)$$

$$h_2(x_1,x_2,\eta) = -\eta + e^{-\eta x_1}(x_2+ \eta)$$

$$h_3(x_1,x_2,\eta) = \eta + e^{\eta x_1}(x_3- \eta)$$

$$\epsilon_1 = \mathrm{sgn}(-x_1)$$

The initial point (x_{10},x_{20},x_{30}) is steered into the origin in the optimal time. This is accomplished by determining the initial value of $\epsilon = \mathrm{sgn}(\sigma)$, and following along the trajectory passing through (x_{10},x_{20},x_{30}) until σ switches sign (having just passed through 0). We then switch ϵ, set t back to 0, set new initial conditions and progress along the new trajectory until σ again switches sign. We repeat this process until the point reaches the origin.

In actuality, the error accumulated from overshooting the switching surfaces prevents us from reaching exactly the origin. In fact, although we should always be able to get to the origin from anywhere in the space in two switchings, we allow the possibility of more switchings if they tend to drive the point closer to the origin. We stop when the chatter caused by proximity to the origin is greater than the distance involved, causing the point to circle the origin endlessly in a limit cycle.

When this system was simulated on an IBM 1620 computer, it was first necessary to produce test points whose optimal times to the origin was known. This we accomplished, in the same manner as in I(a) by moving backwards from the origin, three specified times, arriving finally at our initial point. The control law is only valid for the set

$$|x_2| < 1, \quad |x_3| < 1.$$

When the control law was simulated, we found that although we allowed more than two switchings, it was always the case that the point was closer to the origin after two switchings than three or four. We have therefore reproduced only the data for the first two switchings.

Δt	Point	Time	Time	Distance
		Actual	Computed	
.05	-.30,-.21,-.52	.3, .3, .3	.3, .35,.35	.05
.05	.18, .13, .32	.27,.36,.27	.30,.50,.40	.09
.05	.39, .33, .47	.52,.24,.15	.55,.35,.20	.03
.05	.05,-.02, .16	.00,.25,.30	.05,.35,.35	.05
.02	-.30,-.21,-.52	.3, .3, .3	.30,.32,.33	.02
.01	-.18,-.13,-.32	.27,.36,.27	.27,.38,.29	.01
.-1	-.05, .02,-.16	.00,.25,.30	.01,.27,.31	.01

It will be seen that the smallest distance to the origin after two switches is of the same order of magnitude as the t-increment.

(c) Consider the third order system with one zero and two equal, but opposite in sign, real eigenvalues

$$\dddot{x} - \lambda^2\dot{x} = \epsilon + w(t)$$

where $|\epsilon| \leqq 1$ and $w(t)$ is an external disturbance. The system may be represented in vector form as

$$\dot{y} = \bar{A}y + \bar{a}(\epsilon + w) \qquad (1)$$

or

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \lambda^2 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} (\epsilon + w)$$

In (1) the state variables have a physical significance: $x = y_1$, $\dot{x} = y_2$, and $\ddot{x} = y_3$. Equation (1) may be written in a much simpler mathematical form, however, the state variables lose their physical significance. With the transformation

$$x = Qy$$

where Q is such that

$Q \bar{A} Q^{-1} = A$

and

$Q \bar{a} = a,$

equation (1) becomes

$$\dot{x} = Ax + a(\epsilon + w) \qquad (2)$$

In (2),

$$A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & -\lambda \end{bmatrix} \text{ and } a = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

It should be noted that the region of controllability for the system with a disturbing force is different from that of the system with no disturbances. In Section I(b) the region of controllability of the undisturbed system is the set $|x_2| < \frac{1}{\lambda}$. Now consider the system with external disturbances

$$\dot{x} = Ax + a(\epsilon + w)$$

Further, suppose $|w(t)| \leq M$ on $0 \leq t \leq t_1$, where t_1 is the time taken to reach the origin and $M < 1$ of course. Now the region of controllability is given by the set

$$|x_2| < \frac{1-M}{\lambda}$$

Methods of Simulation

Three methods of simulating system (2) are presented. The first considers using the control law derived in I(b) where $w(t) = 0$. The second is a method of calculating the optimum switching time in the face of some external disturbance. Results are presented in a later section for these two cases. The final method presents an iterative scheme of calculating the initial condition of the adjoint vector in the face of some known external disturbance.

(i) First Method

If a state vector in space is considered and it is desired to move that vector to some desired position by use of the control law developed in I(b) in the face of some external disturbance, it can be assumed that the time will not be the optimum time of the no disturbance case. The effect of the disturbance on the system trajectory will vary from a decrease in time to travel to the origin to an increase as compared to the no disturbance case elapsed time. The reason for this is that as soon as there is a disturbance in the system, a new system is being dealt with and the control law no longer applies.

The time may well be shorter than for the no disturbance optimum case, but it is not optimal for the new case. From the practical point of view, however, since it is unlikely that all disturbances will ever be accurately predicted it is desirable to know how the control law will behave in the face of disturbances.

It was decided to test the control law by simulating constant force disturbances, up to a magnitude of ± 25 percent of ϵ in equation (2) with $\lambda = 1$. The results of this simulation are shown in Figures 7,8,9. The magnitude of the disturbance applied is scaled on the abscissa, both positive and negative, with the center the no disturbance case. The ordinate is scaled in time for the vector to reach the origin, with all times normalized to the no disturbance optimum time.

(ii) Second Method

The second method is strictly a brute force method of calculating the switching times of the system with the external disturbance present. The solution of (2) is given by

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-As}a(\epsilon + w)ds \qquad (3)$$

Since we are interested in the control $\epsilon = \pm 1$ which steers $x(t)$ from $x(0)$ to the origin in mimimum time and the control function is unique, then any $\epsilon = \pm 1$, which steers $x(t)$ to the origin, is the optimum control function. With this in mind (3) may be rewritten as

$$-x(t) = \int_0^{t_f} e^{-As}a(\epsilon + w(s))ds \tag{4}$$

where t_f is the time where $x(t) = 0$. Further, the optimal unique function is $\epsilon(t) = \pm 1$ and there is at most 2 switches in the sign of ϵ as $x(t)$ is steered from $x(0)$ to the origin. Thus (4) becomes

$$-x(0) = U[\int_0^{t_1} e^{-As}ads = \int_1^{t_2} e^{-As}ads + \int_2^{t_f} e^{-As}ads] + \int^{t_f} e^{-As}aw(s)ds \tag{5}$$

where U is the initial sign of $\epsilon(t)$ and where t_1 and t_2 are the first and second switching times respectively. Thus, if t_1, t_2, and t_f, the initial value of ϵ, and the disturbing function $w(t)$ are known, then the problem is solved.

In order to write (5) in detail, one must evaluate e^{At} from the series

$$e^{At} = I + At + A^2 \frac{t^2}{2} + \ldots$$

Thus

$$e^{-As} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & e^{-\lambda s} & 0 \\ 0 & 0 & e^{\lambda s} \end{bmatrix}$$

and

$$e^{-As} a = \begin{bmatrix} 1 \\ e^{-\lambda s} \\ e^{\lambda s} \end{bmatrix}$$

Written out in full, equation (5) becomes

$$-x_1(0) = U[\int_0^{t_1} ds - \int_{t_1}^{t_2} ds + \int_{t_2}^{t_f} ds] + \int_0^{t_f} w(s)ds$$

$$-x_2(0) = U[\int_0^{t_1} e^{-\lambda s} ds - \int_{t_1}^{t_2} e^{-\lambda s} ds + \int_{t_2}^{t_f} e^{-\lambda s} ds] + \int_0^{t_f} w(s)e^{-\lambda s} ds$$

$$-x_3(0) = U[\int_0^{t_1} e^{\lambda s} ds - \int_{t_1}^{t_2} e^{\lambda s} ds + \int_{t_2}^{t_f} e^{\lambda s} ds] + \int_0^{t_f} w(s)e^{\lambda s} ds \qquad (6)$$

The problem of solving this set of equations for the times t_1, t_2, t_f becomes intolerable unless some suitable constraint is put on the sorts of disturbances one expects to encounter. The solution is simple for a constant disturbance $w(t) = C$ but this is an unrealistic assumption. A much stronger solution would result from allowing $w(t)$ to be of the form

$$w(t) = b-ce^{-at}$$

which is shown in figure 1.

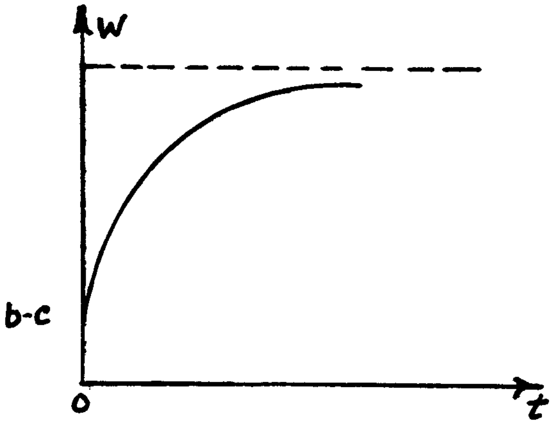


FIGURE 1

But this wind makes equation (6) impossible to solve by any algebraic methods. Hence the compromise assumption is made, namely,

$$w(t) = \begin{cases} b-ce^{-at}, & t \leqq d \leqq t_f \\ b \quad , & t > d \end{cases} \tag{7}$$

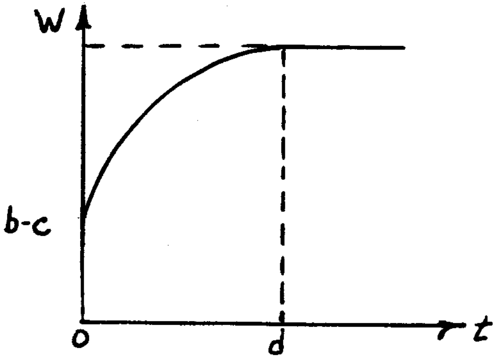


FIGURE 2

Where d is known in advance we have

$$-x_1(0) = U[2t_1 - 2t_2 + t_f] + bd + \frac{c}{a}(e^{-ad}-1) + bt_f - bd$$

$$-x_2(0) = -\frac{U}{\lambda}[2e^{\lambda t_1} -1 -2e^{-\lambda t_2} + e^{-\lambda t_f}] - \frac{b}{\lambda}(e^{-\lambda t_f}-1) + \frac{c}{\lambda + a}(e^{(-\lambda-a)d}-1)$$

$$-x_3(0) = \frac{U}{\lambda}[2e^{\lambda t_1}-1 -2e^{\lambda t_2} + e^{\lambda t_f}] + \frac{b}{\lambda}(e^{\lambda t_f}-1) - \frac{c}{\lambda - a}(e^{(\lambda-a)d}-1)$$

(8)

$$2t_1 - 2t_2 + t_f(1 + bu) + U[\frac{c}{a}(e^{-ad}-1) + x_1(0)] = 0$$

$$2e^{-\lambda t_1} - 2e^{-\lambda t_2} + e^{-\lambda t_f}(1 + bu) - \lambda U[b + \frac{c}{x + a}(e^{(-\lambda - a)d}-1) + x_2(0)]-1 = 0$$

$$2e^{\lambda t_1} - 2e^{\lambda t_2} + e^{\lambda t_f}(1 + bU) + \lambda U[-b - \frac{c}{\lambda - a}(e^{(\lambda - a)d}-1) + x_3(0)]-1 = 0 \tag{9}$$

or

$$2t_1 - 2t_2 + ft_f + R_1 = 0$$

$$2e^{-\lambda t_1} - 2e^{-\lambda t_2} + fe^{-\lambda t_f} + R_2 = 0$$

$$2e^{\lambda t_1} - 2e^{\lambda t_2} + fe^{\lambda t_3} + R_3 = 0 \tag{10}$$

where

$$R_1 = U[\frac{c}{a}(e^{-ad} -1) + x_1(0)]$$

$$R_2 = -\lambda U[b + \frac{c}{\lambda + a}(e^{(-\lambda - a)d} -1) + x_2(0)]-1$$

$$R_3 = \lambda U[-b - \frac{c}{\lambda - a}(e^{(\lambda - a)d} - 1) + x_3(0)]-1$$

$$f = 1 + bU$$

This set of equations may be solved easily.

$$t_2 = t_1 + \frac{1}{2}(ft_f + R_1)$$

$$2e^{-\lambda t_1} - 2e^{\lambda[-t_1 - \frac{1}{2}(ft_f + R_1)]} + fe^{-\lambda t_3} + R_2 = 0 \qquad (11)$$

$$2e^{\lambda t_1} - 2e^{\lambda[t_1 + \frac{1}{2}(ft_f + R_1)]} + fe^{\lambda t_3} + R_3 = 0 \qquad (12)$$

From (12) we get

$$e^{\lambda t_1} = (fe^{\lambda t_f} + R_3)/2(e^{\frac{1}{2}\lambda(ft_f + R_1)} - 1)$$

By substituting this expression for $e^{\lambda t_1}$ into equation (11) and incrementing t_f from d (since it is known that t_f is no smaller than d), one need only observe the value of t_f at which equation (11) changes sign. If at this point the coordinates

$$0 \leqq t_1 \leqq t_2 \leqq t_f$$

are satisfied, then it is guaranteed that t_1, t_2, t_f are the optimal switching times.

This control scheme was simulated on an IBM 1620 Computer with results reproduced later. There are two difficulties involved in choosing the initial conditions for a test: choosing the correct initial value U of ϵ , and choosing a d which is not unrealistically small but which is nevertheless smaller than a readonable prediction for t_f. If D is the distance from the point $x(0)$ to the origin, then the fastest time in which a point on the D sphere could possibly reach the origin is about $\frac{1}{2}$ D - its path being of course, the switching curve through the origin. Hence it is safe to assign d any value less than $\frac{1}{2}$ D. The matter of choosing the correct initial ϵ is more difficult and more serious. An equation for the switching surface σ described in I(b) will give the initial value, but use of this device seems self-defeating. Lacking that, the choice is arbitrary.

(iii) Third Method

An Iterative Scheme for Calculating the Control Function

Rather than calculate the switching surfaces for a system, or calculate the switching times, one might consider a method of solving for the initial condition of the adjoint vector to the system. The following is an example of the latter procedure.

It is well known that the time optimal control function $\epsilon(t)$ for the system

$$x = Ax + a\epsilon, \quad |\epsilon| \leqq 1$$

is given by

$$e(t) = \text{sgn}[\eta(t)\cdot a]$$

where $\eta(t)$ is the solution of the adjoint system

$$\mathring{\eta} = -A'\eta$$

with $\eta(0) = \eta_o$. The solution of this adjoint system is

$$\eta(t) = e^{-A't}\eta_o$$

Thus the optimal control function is given by

$$\epsilon(t) = \text{sgn}[\eta_o\cdot e^{-At}a].$$

The time optimal problem is solved is the initial condition of the adjoint vector $\eta(0) = \eta_o$ is known. The following method is an iterative scheme to determine the value of η_o for a given x_o. The method is due to Neustadt.* The method will be described first for an autonomous system without external disturbances and then

* Neustadt, L.W., Synthesizing Time Optimal Controls. Journal of Mathematical Analysis and Applications, vol. 1, no. 3, December 1960.

extended to the nonautonomous system with deterministic wind disturbances.

I. Consider the first case, that is, the system governed by

$$\dot{x} = Ax + a\epsilon, \quad |\epsilon| \leq 1 \tag{13}$$

Consider the set of attainability $C(t)$ which is the set of initial conditions x_o from which the origin can be reached in time t with control $\epsilon(t)$. Neustadt proves that $C(t)$ is closed, convex, and nonempty. If the system is normal, then the boundary of $C(t)$, $\partial C(t)$, contains no straight line segments. The system is said to be normal if the function $\eta_o \cdot e^{-At}a$ has a countable number of zeros, that is $\epsilon(t)$ is defined almost everywhere. Further, if x_o lies on $\partial C(t)$, then an extremal control

$$\epsilon(t) = \mathrm{sgn}[\eta_o \cdot e^{-At}a], \tag{14}$$

is required to reach the origin, where η_o is the exterior normal to $C(t)$ at $-x_o$. $C(t) \supset C(t')$ if $t > t'$ and $C(t)$ grows continuously with t.

The solution of (13) using the optimum control function (14) is given by

$$x(t) = e^{At}x_o + e^{At}\int_0^t e^{-As} a \operatorname{sgn}[\eta_o \cdot e^{-As}a]ds \tag{15}$$

Let t be the minimum time to reach the origin. Then (15) becomes

$$-x_o = \int_0^t e^{-As} a \operatorname{sgn}[\eta_o \cdot e^{-As}a]ds \tag{16}$$

Define

$$z(t,\eta_o) = \int_0^t e^{-As} a \operatorname{sgn}[\eta_o \cdot e^{-As} a]ds$$

Surely for any $\xi \epsilon C(t)$ and $\xi \neq z(t,\eta_o)$, $\eta_o \cdot z(t,\eta_o) > \eta_o \cdot \xi$.

Also

$$\eta_o \cdot z(t,\eta_o) = \int_0^t \eta_o \cdot e^{-As} a \operatorname{sgn}[\eta_o \cdot e^{-As} a]ds = \int_0^t |\eta_o \cdot e^{-As}a|ds > 0$$

Thus $\eta_o \cdot z(t,\eta_o)$ is a monotone increasing function of t for fixed η_o. Figure 3 shows the geometrical relation of the vector z and η_o for the correct value of η_o. Note that for $\|\eta_o\| = 1$, the line $op = \eta_o \cdot \xi < oq = \eta_o \cdot z(t,\eta_o)$ for any $\xi \epsilon C(t)$ and $\xi \neq z(t,\eta_o)$.

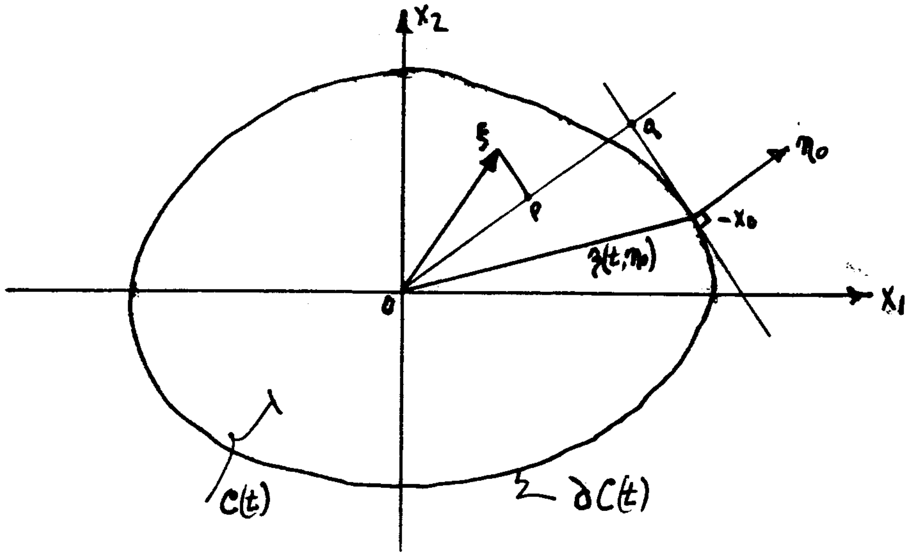


FIGURE 3

Now form the scalar function

$$f(t,\eta_o,x_o) = \eta_o\cdot[z(t,\eta_o) + x_o] = \int_0^t |\eta_o\cdot e^{-As}a|ds + \eta_o\cdot x_o \qquad (17)$$

which is continuous in t,η_o, and x_o and, for fixed η_o and x_o, is a strictly monotone increasing function of t.

Since η_o is the unknown that we are going to find through an iterative scheme, an initial value of η_o is chosen such that

$$\eta_o^{(1)}\cdot x_o = f(0,\eta_o^{(1)}, x_o) < 0.$$

A common choice for $\eta_o^{(1)}$ is

$$\eta_o^{(1)} = - \frac{x_o}{\|x_o\|} .$$

If $C(t)$ was a hypersphere then this initial guess at η_o would indeed be the correct value. With this initial choice $\eta_o^{(1)}$, let t increase until $f(t,\eta_o^{(1)}, x_o) = 0$ and denote this value of t by $t^{(1)}$. The geometrical picture of the above statement is shown in Figure 4. If $z(t^{(1)}, \eta_o^{(1)}) \neq -x_o$ then the choice $\eta_o^{(1)}$ was incorrect. A series of corrections to the η_o vector is necessary in order to converge to the correct value. Neustadt suggests the steepest descent method.

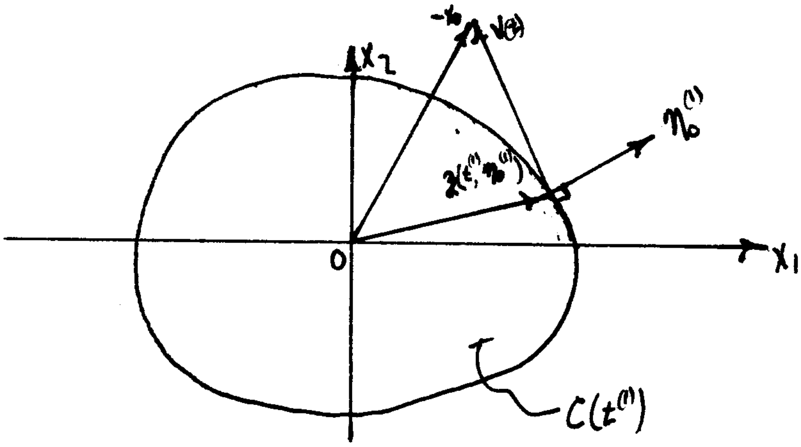


FIGURE 4

The correction in $\eta_o^{(1)}$ should lie along the "error vector" $v(t^{(1)})$ where

[redacted]

[redacted]

$$v(t^{(1)}) = -x_o - z(t^{(1)}, \eta_o).$$

Thus the corrected value of $\eta_o^{(1)}$, called $\eta_o^{(2)}$, is given by

$$\eta_o^{(2)} = \eta_o^{(1)} - [k\, x_o + z(t^{(1)}, \eta_o^{(1)}] \qquad (18)$$

A new $f(t, \eta_o^{(2)}, x_o)$ is formed and a $t^{(2)}$ is calculated. Again, if $z(^{(2)}, \eta_o^{(2)}) \neq -x_o$ then the procedure is continued. The cycle is repeated until $z(t^{(i)}, \eta_o^{(i)})$ is within a small distance δ of $-x_o$.

The value of k in (18) will affect the rate of convergence of $\eta_o^{(i)}$ to the correct value. Paiewansky* used this procedure on the second order system,

$$\ddot{x} + 0.1x = \epsilon(t),$$

and states that increasing k decreases the number of iterations required to reach the optimum η_o. Further increase in k results in

* Paiewansky, Bernard H., "The Synthesis of Optimum Controllers" presented at the Optimum System Synthesis Conference, September 1962. The work was sponsored partly by the Wright-Patterson Air Force Base, Aeronautical Systems Division, Contract Number AF 33(677)-7781.

large and undesirable oscillations about the correct value.

In this procedure, we must solve the equation

$$f(t,\eta_o,x_o) = 0.$$

This defines a new function

$$t = T(\eta_o,x_o).$$

Neustadt shows the correction vector to $\eta_o^{(i)}$, namely, $-[x_o + z(t^{(i)}, \eta_o^{(i)})]$ is indeed proportional to grad. T, that is

$$\nabla T = - \frac{[x_o + z(t,\eta_o)]}{|\eta_o \cdot e^{-At}a|}$$

As a trivial example of this method, consider the system

$$\ddot{x} = \epsilon$$

or

$$\dot{x}_1 = z_2$$

$$\dot{x}_2 = \epsilon \qquad (19)$$

The solution of (19) is given by

$$\begin{bmatrix} x_1(t) \\ x_2(t)\end{bmatrix} = \begin{bmatrix}1 & t\\ 0 & 1\end{bmatrix}\begin{bmatrix}x_1(0)\\ x_2(0)\end{bmatrix} + \begin{bmatrix}1 & t\\ 0 & 1\end{bmatrix}\int_0^t \begin{bmatrix}1 & -S\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} \text{sgn} \begin{bmatrix}\eta_1(0)\\ \eta_2(0)\end{bmatrix} \cdot \begin{bmatrix}1 & -S\\ 0 & 1\end{bmatrix}\begin{bmatrix}0\\ 1\end{bmatrix} ds$$

Equation (16) becomes

$$\begin{bmatrix} -x_1(0) \\ -x_2(0)\end{bmatrix} = \int_0^t \begin{bmatrix} -s \\ 1\end{bmatrix} \text{sgn}[\eta_2(0) - s\eta_1(0)]ds$$

Consider $x_o = (-1,0)$ then

$$\eta_o^{(1)} = \frac{-x_o}{\|x_o\|} = \binom{1}{0}$$

Forming $f(t, \eta_o^{(1)}, x_o)$ from (17), one has that

$$f(t, \eta_o^{(1)}, x_o) = \frac{t^2}{2} - 1$$

Setting $f(t, \eta_o^{(1)}, x_o) = 0$ yields $t^{(1)} = \sqrt{2}$

Now using (18) to calculate a new η_o

$$\eta_o^{(2)} = \eta_o^{(1)} - k[x_o - z(t^{(1)}, \eta_o^{(1)}] = \binom{1}{1}$$

where k is chosen as $\frac{1}{\sqrt{2}}$.

$$f(t,\eta_o^{(2)}, x_o) = \frac{t^2}{2} - t, \quad t^{(2)} = 2.$$

Forming $z(t^{(2)}, \eta_o^{(2)})$ one sees that

$$z(t^{(2)}, \eta_o^{(2)}) = -x_o \text{ .}$$

Thus $\eta_o^{(2)}$ is the correct value. The optimum control function is given by

$$\epsilon(t) = \text{sgn}\begin{bmatrix} \eta_1^{(2)}(0) \\ \eta_2^{(2)}(0) \end{bmatrix} \cdot \begin{bmatrix} 1 & -t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \text{sgn}[1-t]$$

II. Consider the second case, that is, the nonautonomous normal systems with external disturbances,

$$\dot{x} = A(t)x + a(t)\epsilon + w(t) \tag{20}$$

where $w(t)$ is a known vector function. For our problem, $A(t)$ and $a(t)$ will be constants although this more general system is presented for completeness. The solution of (20) is given by

$$x(t) = \Phi(t)x_o + \int_{t_o}^{t} \Phi(t)\Phi^{-1}(s)a(s)\epsilon(s)ds + \int_{t_o}^{t} \Phi(t)\Phi^{-1}(s)w(s)ds \tag{21}$$

where $\Phi(t)$ is the fundamental solution matrix satisfying the equation

$$\dot{\Phi}(t) = A(t)\Phi(t), \ \Phi(t_o) = I$$

of course, $x_o = x(t_o)$.

The time optimal regulator problem consists in choosing an $\epsilon(t)$ on $t_o \leqq t \leqq T$ such that $x(T)$ is zero for minimum T. With this in mind, (21) may be written as

$$-x_o - \int_{t_o}^{T} \Phi^{-1}(s)w(s)ds = \int_{t_o}^{T} \Phi^{-1}(s)a(s)\epsilon(s)ds.$$

Let

$$w(t) = -x_o - \int_{t_o}^{t} \Phi^{-1}(s)w(s)ds$$

Recall that the optimum allowable control function is of the form

$$\epsilon(t) = \text{sgn}[\eta_o \cdot \Phi^{-1}(t)a(t)]$$

The regulator problem now consists in choosing an η_o such that

$$\int_{t_o}^{T} \Phi^{-1}(s)a(s)\text{sgn}[\eta_o\Phi^{-1}(s)a(s)]ds = \omega(t)$$

for minimum T.

Let $C(t)$ be the set of all points $\int_{t_o}^{T} \Phi^{-1}(s)a(s)\epsilon(s)ds$, which can be reached from the origin using all allowable controls in time t. As before, $C(t)$ is a closed convex set. As in the constant coefficient case, let

$$z(t,\eta_o) = \int_{t_o}^{t} \Phi^{-1}(s)a(s)\text{sgn}[\eta_o\cdot\Phi^{-1}(s)a(s)]ds \ldots$$

Since $z(t,\eta_o)$ is an extremal response, it lies on $\partial C(t)$ and η_o is the exterior normal to $C(t)$ at $z(t,\eta_o)$. As in the previous case, we define a new scalar function

$$f(t,\eta_o,\omega) = \eta_o\cdot[z(t,\eta_o) - \omega(t)]$$

Choose $\eta_o^{(1)}$ such that $\eta_o^{(1)}\cdot\omega(t_o) > 0$, a common choice being

$$\eta_o^{(1)} = \frac{\omega(t_o)}{\|\omega(t_o)\|}$$

Now

$$f(t_o,\eta_o^{(1)}, \omega) = -\eta_o^{(1)}\cdot\omega(t_o) < 0$$

Let t increase from t_o until $f(t,\eta_o^{(1)},\omega) = 0$ and denote this value by $t^{(1)}$. As in the constant coefficient case choose a new η_o as follows

$$\eta_o^{(2)} = \eta_o^{(1)} + kv(t^{(1)}) = \eta_o^{(1)} + k[\omega(t^{(1)}) - z(t^{(1)}, \eta_o^{(1)})]$$

This iteration process is continued until $\eta_o^{(i+1)} = \eta_o^{(i)}$. This final $\eta_o^{(i)}$ is the correct value. Figure (5) illustrates the geometrical picture of the above statements.

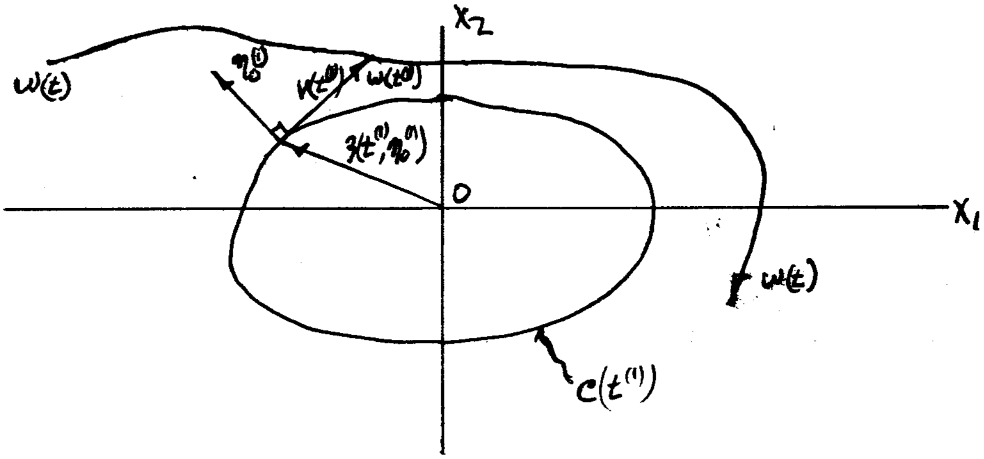


FIGURE 5

(iv) Comparison of Methods

The control laws were tested under various sorts of disturbances. First, the switching surface control, in which the switching surfaces for no disturbance were computed in advance and used to determine switching points, was employed to drive points into the origin against small constant constraints. Next the method of computing switching times directly was used. With this method the optimal times for the wind-blown system were found. These times are compared, in a series of graphs below, with the non-optimal times for the same plants under the same disturbances.

Finally, the time-computing method was used to study the behavior of the plants under variable constraint conditions. In all cases, as stated above it is required that the disturbance became constant before the plant reaches the origin.

The formula for the kind of disturbances in the second method is

$$w(t) = \begin{cases} h-ce^{-at}, & t \leq d < t_f \\ b, & t > d \end{cases}$$

This typical wind is shown in Figure 6.

<u>OPTIMAL CONTROL UNDER CONSTRAINTS - METHOD 2</u>

Δt		a	b	c	Time	Distance
"	(-.30, -.21, -.52)	6.0	.0	0	.9	.0001
"	"	"	.05	.8	1.1	.0058
"	"	"	.15	"	1.05	.0049
"	(.30, .02, -.16)	"	0	0	.9	.0001
"	"	"	.1	0	.95	.0068
"	(-.05, .02, -.16)	"	0	0	.55	.0016
"	"	10.0	.05	.6	.60	.0064
"	"	"	-.05	"	.60	.0038
"	"	"	.15	"	.60	.0084
"	"	"	-.15	"	.60	.0003
"	"	"	.25	"	.60	.0097
"	"	"	-.25	"	.65	.0090
"	(-.39, -.33, -.47)	"	0	0	.90	.0016
"	"	5.0	.15	.1	.75	.0011
"	(.18, .13, .32)	10.0	0	.0	.90	.0019
"	"	5.0	.15	.8	1.05	.0037
"	"	6.0	-.15	.8	1.20	.0114
"	"	"	.25	"	1.0	.0007
"	"	"	.15	0	.90	.0080
"	"	"	-.15	0	.95	.0061
"	"	"	0	.8	1.1	.0079

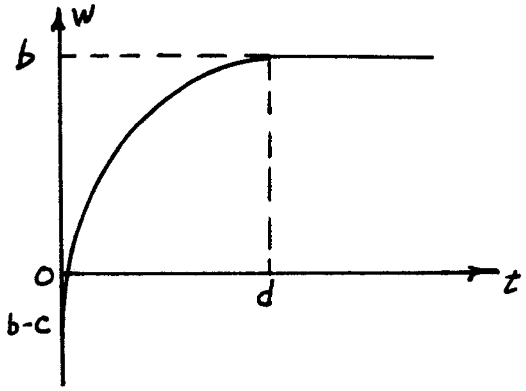


FIGURE 6

Several things are noteworthy about this method. First of all, although the running procedure is essentially the same as that of method 1 — hunting for a change of sign in a polynomial (see "Gilchrist Control Law" Fortran program in Appendix) — the second method gives much greater accuracy with respect to the time increment. The final distance to the origin is less than one tenth the distance found in the other procedure, and the times found are ten to forty percent more accurate. Such a gain in accuracy, however, is more than offset by the unfeasible necessity of knowing the wind velocity in advance. Thus, while method 2 has many theoretical advantages, its information prerequisites make it somewhat impractical.

A comparison of the computed times to the origin by methods 1 and 2 follows. In method 1 no allowance was made for the wind, which merely disturbed the normal path to the origin. In method 2, the optimal paths were computed with the wind disturbance taken into account. The times have been normalized with respect to the (known) optimal times. Refining the t-increment would have the double effect of smoothing both curves, and moving the two points on the t-axis closed to the correct position $D = 0$; $t = 1$.

The following data was obtained from various values of a,b,c, d being set arbitrarily to $3/a$; and λ, as, throughout, equal to $+ 1$.

The simulation using method 1 indicates that the control law is workable in the face of disturbances which, in effect, cause the state vector to move according to a non-optimal control law. A rough estimate from Figures 7,8, and 9 show that the greatest increase in time is approximately 40 percent, although it must be admitted that part of this increase may be due to inaccuracies in the simulation (too large a time increment). It is difficult to estimate whether this increase in time would be good, bad or indifferent in a practical system. A general indication of its practicality can be obtained by referring to figures 7,8, and 9 which also illustrate simulation carried out by assuming advance knowledge of the disturbances and adjusting the control law accordingly.

Appendix contains the two FORTRAN programs used to execute these two tests, method 1 (CONTROL LAW III WITH STEP NOISE) and method 2 (GILCHRIST CONTROL LAW).

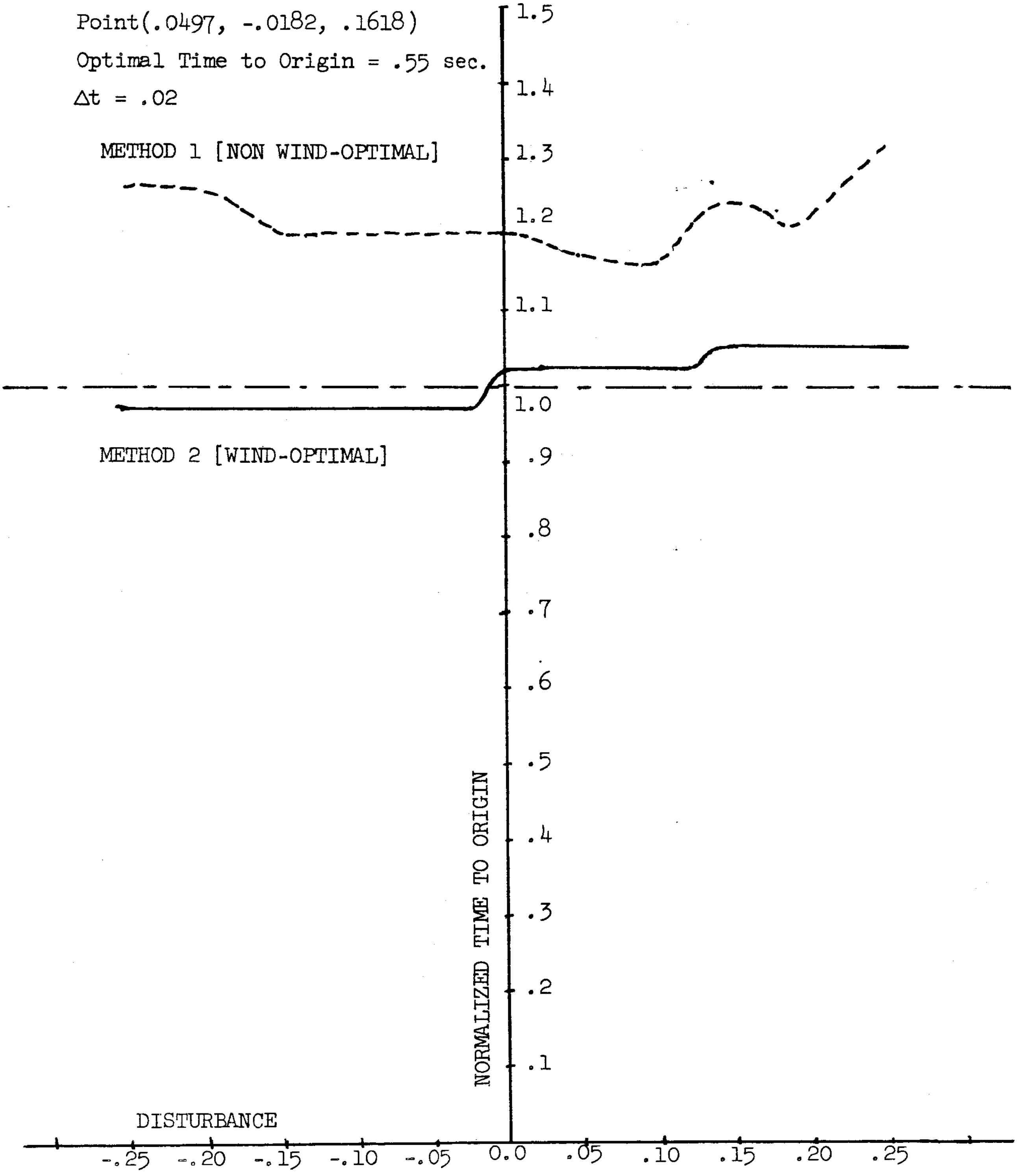

Point(.0497, -.0182, .1618)
Optimal Time to Origin = .55 sec.
Δt = .02
METHOD 1 [NON WIND-OPTIMAL]
METHOD 2 [WIND-OPTIMAL]
NORMALIZED TIME TO ORIGIN
DISTURBANCE
1.5
1.4
1.3
1.2
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
-.25
-.20
-.15
-.10
-.05
0.0
.05
.10
.15
.20
.25

FIGURE 7

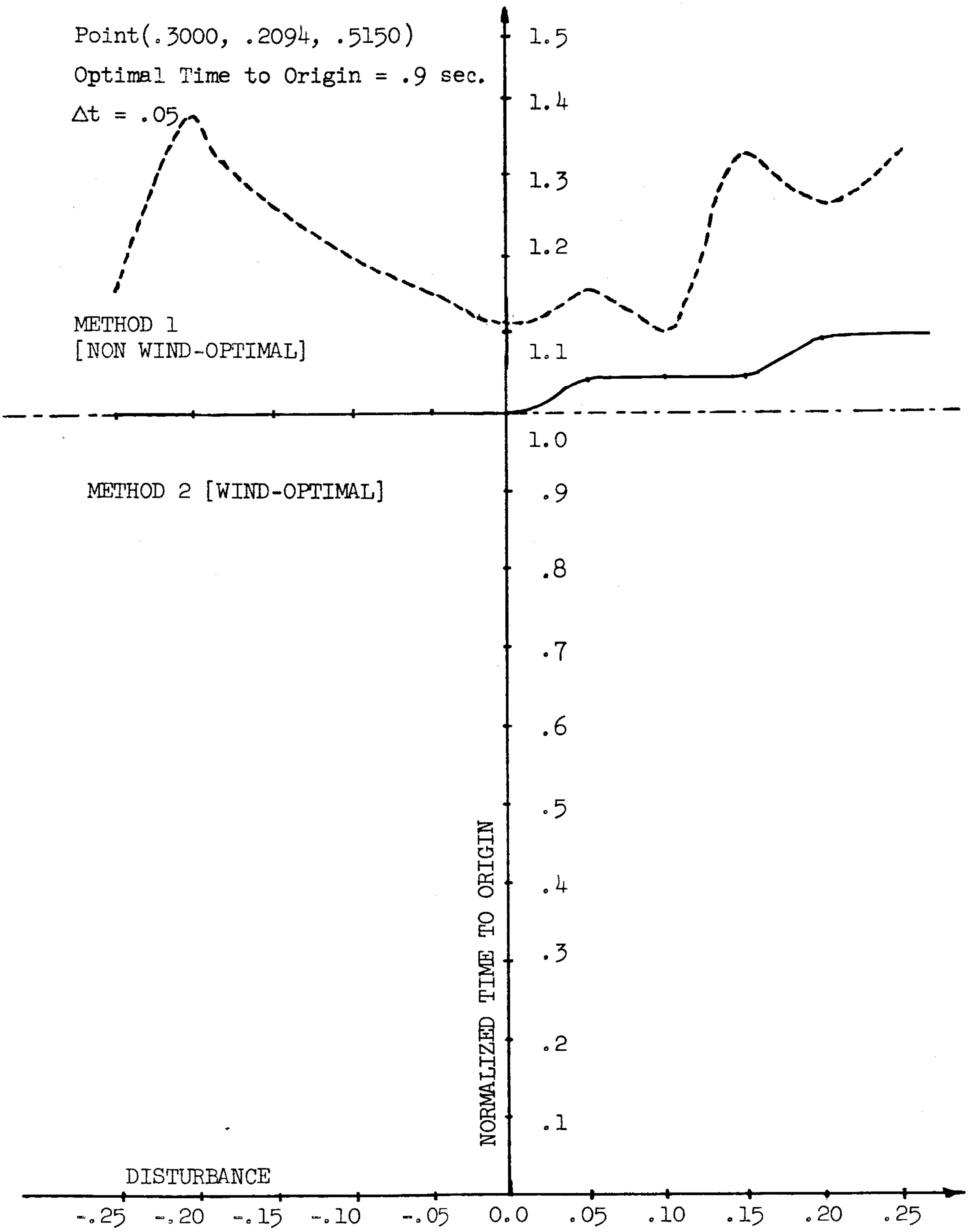

Point(.3000, .2094, .5150)
Optimal Time to Origin = .9 sec.
Δt = .05
METHOD 1
[NON WIND-OPTIMAL]
METHOD 2 [WIND-OPTIMAL]
NORMALIZED TIME TO ORIGIN
DISTURBANCE
1.5
1.4
1.3
1.2
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
-.25
-.20
-.15
-.10
-.05
0.0
.05
.10
.15
.20
.25

FIGURE 8

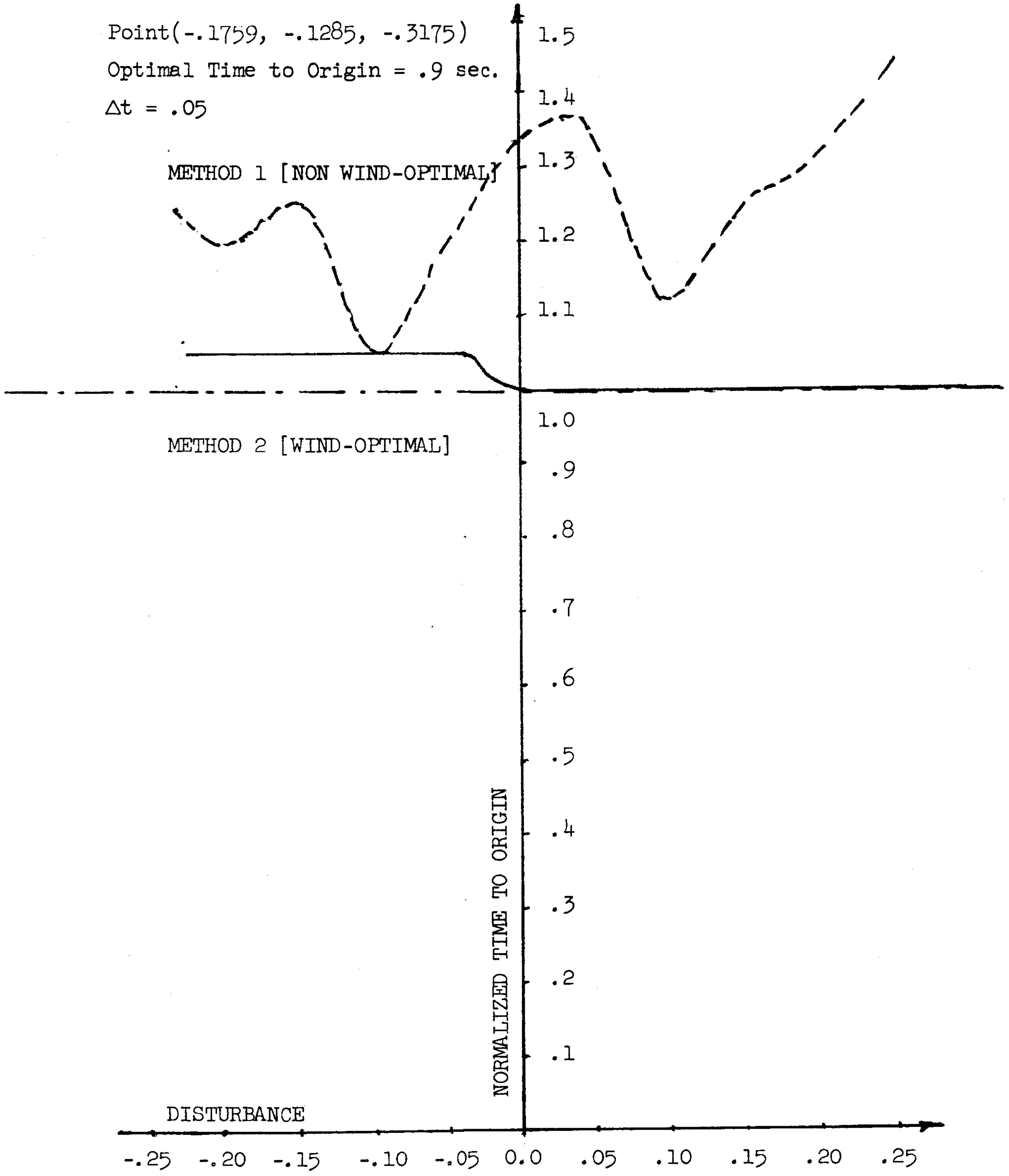

Point(-.1759, -.1285, -.3175)
Optimal Time to Origin = .9 sec.
Δt = .05
METHOD 1 [NON WIND-OPTIMAL]
METHOD 2 [WIND-OPTIMAL]
1.5
1.4
1.3
1.2
1.1
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
NORMALIZED TIME TO ORIGIN
DISTURBANCE
-.25
-.20
-.15
-.10
-.05
0.0
.05
.10
.15
.20
.25

FIGURE 9

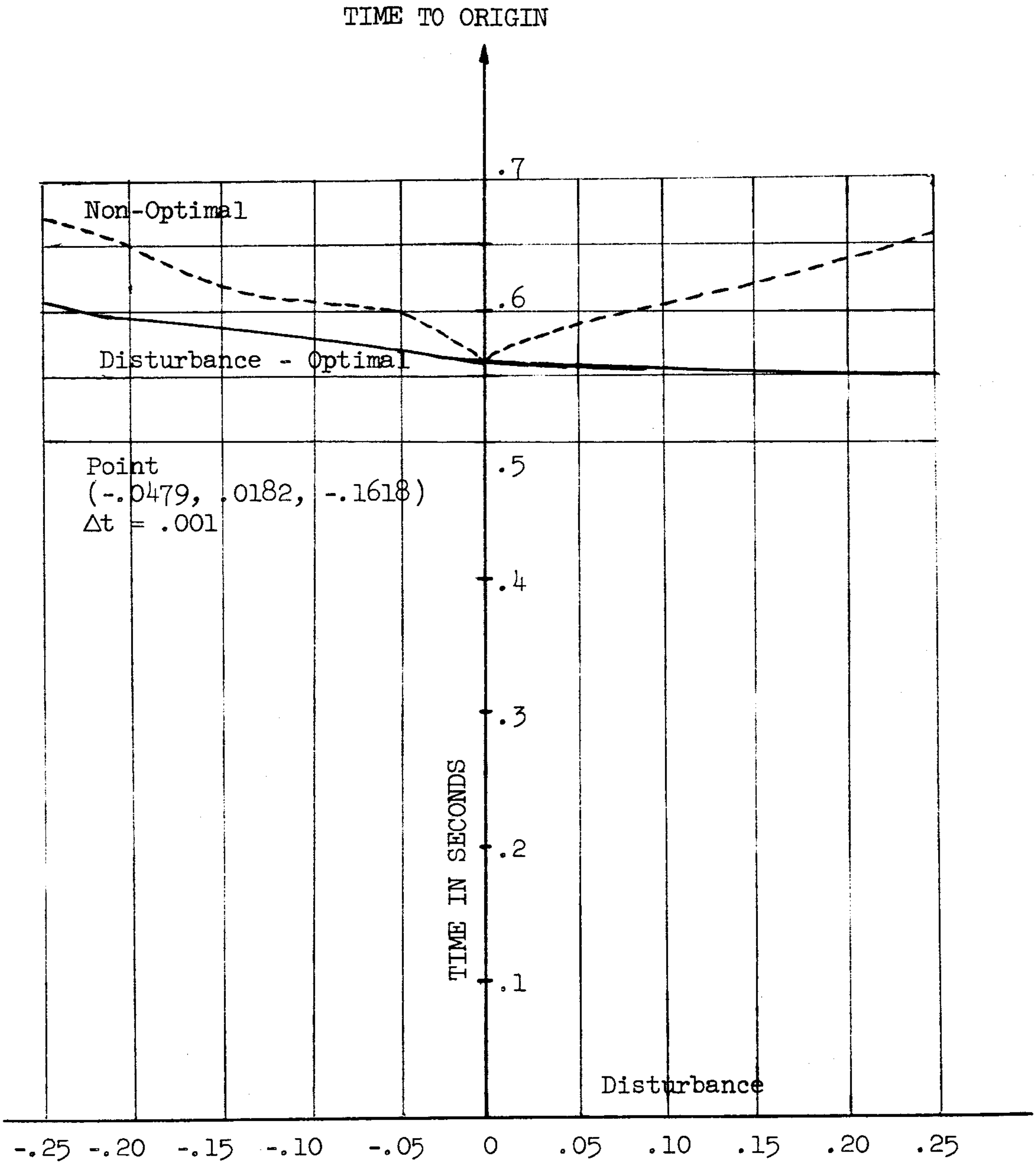

TIME TO ORIGIN
.7
.6
.5
.4
.3
.2
.1
TIME IN SECONDS
Non-Optimal
Disturbance - Optimal
Point
(-.0479, .0182, -.1618)
Δt = .001
Disturbance
-.25
-.20
-.15
-.10
-.05
0
.05
.10
.15
.20
.25

FIGURE 10

Synthesizing Performance Under Disturbances

The system

$$\dddot{x} - \lambda^2 \dot{x} = \epsilon$$

can be controlled optimally by a number of different schematizations of the same control law. Some of these methods involve discovering expressions for the switching surfaces, precomputing the switching times, or treating the switching time equations themselves as formulas for the switching surfaces. The first two methods were described extensively in the previous sections I(b) and I(c). The last method, though less highly developed than the others, has some merit and is described below.

The solution to the (transformed) system, which is affected by some external disturbance w

$$\dot{x} = Ax + b(\epsilon + w) \tag{22}$$

is given by

$$x(t) = e^{At}x(0) + e^{At} \int_0^t e^{-As} b(\epsilon + w) ds \tag{23}$$

Since it is desired to reach the origin in time t, $x(t) = 0$. Hence

$$-x(0) = \int_0^t e^{-As} b(\epsilon + w) ds, \qquad (24)$$

where

$$A = \begin{vmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix}$$

$$b = \begin{vmatrix} 1 \\ 1 \\ 1 \end{vmatrix}$$

$$e^{-As} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & e^{-\lambda s} & 0 \\ 0 & 0 & e^{\lambda s} \end{vmatrix}, \quad \text{and}$$

$$e^{-As} b = \begin{vmatrix} 1 \\ e^{-\lambda s} \\ e^{\lambda s} \end{vmatrix}$$

The three equations comprising equation (24) become

$$-x_1(0) = \int_0^t (\epsilon + w)ds$$

$$-x_2(0) = \int_0^t (e^{-\lambda s}(\epsilon + w)ds$$

$$-x_3(0) = \int_0^t e^{\lambda s}(\epsilon + w)ds \qquad (25)$$

Let us make two assumptions: first, that the point $x(0)$ is on the lowest order switching surface*, S_1, second, that the disturbance $w(t)$ is a constant w. Then exactly one switching at time t_2 will be required to reach the origin.

Let U designate the initial value of ϵ. Since we assume that $x(0)$ is on a switching surface S_1, $\epsilon = -U$ on S_1.

$$-x_1(0) = -U[2t_2 - t_3] + wt_3$$

$$-x_2(0) = U[2e^{\lambda t_2} - e^{-\lambda t_3} - 1] - \frac{w}{\lambda}(e^{-\lambda t_3} - 1)$$

$$-x_3(0) = -U[2e^{\lambda t_2} - e^{\lambda t_3} - 1] + \frac{w}{\lambda}(e^{\lambda t_3} - 1).$$

* By lowest order switching surface we understand the surface at which one switches for the first time - assuming that the initial point $x(0)$ was not on a switching surface. Subsequent surfaces are high-order.

Assume $\lambda = 1$.

Let $R_1 = -Ux_1(0)$

$R_2 = U(x_2(0) + w) - 1$

$R_3 = -U(x_3(0) - w) - 1$

$G = 1 + Uw$

Then

$$2t_2 - Gt_3 + R_1 = 0 \qquad a$$

$$2e^{-t_2} - Ge^{-t_3} + R_2 = 0 \qquad b$$

$$2e^{t_2} - Ge^{t_3} + R_3 = 0 \qquad c \qquad (26)$$

From (26b)

$$Ge^{-t_3} = 2e^{-t_2} + R_2$$

$$e^{-t_3} = (2e^{-t_2} + R_2)/G$$

$$e^{t_3} = G/(2e^{-t_2} + R_2)$$

From (26c)

$$2e^{t_2} - G^2/(2e^{-t_2} + R_2) + R_3 = 0$$

$$4 + 2R_2e^{t_2} - G^2 + 2R_3e^{-t_2} + R_2R_3 = 0$$

$$2R_2e^{2t_2} + e^{t_2}(4-G^2 + R_2R_3) + 2R_3 = 0$$

$$e^{t_2} = (-B-B^2-4AC)/2A \qquad (27)$$

Where $A = 2R_2$

$$B = 4-G^2 + R_2R_3$$

$$C = 2R_3$$

$$e^{t_3} = (2e^{t_2} + R_3)/G \qquad (28)$$

The assumed condition, that $x(0)$ be on a switching surface, is true if equation (26a) is satisfied for these values t_2, t_3. The function

$$F(x) = 2t_2 - Gt_3 + R_1 \qquad (29)$$

must therefore be zero if $x(0)$ is on a switching surface. The converse condition $-x(0)$ on S_1 if F is 0 will also hold if, in addition, the inequalities

$$t_3 \geqq t_2 \geqq 0 \qquad (30)$$

hold. Thus we have

$F(x) = 0$ and $t_3 \geq t_2 \geq)$ <=> is on S_1

These conditions then are necessary and sufficient for x to be on a switching surface, of course, if x happens to lie initially on a high order switching surface (in the three-dimensional case, this means that x lies on the trajectory through the origin), $F(x) = 0$, and either $t_2 = t_3$, or $t_2 = 0$, depending on the sign of U. In addition, F is continuous in x, so the magnitude of F provides a measure of how far x is from S_1. By evaluating $F(x)$ as t_1 is incremented from 0, one need only observe the point at which $F(x)$ changes sign, and if the inequalities are satisfied there, x will have just passed through a switching surface. Since we have taken care not to land exactly <u>on</u> the switching surface, the process may be begun again.

The disadvantages of this method are that the sign of $F(x)$ gives no clue as to the sign of ϵ, and that there are large regions on which F cannot be evaluated due to the discriminant's being negative (although it must be non-negative on a switching surface). However, these are not serious difficulties, especially the discriminant problem, because the discriminant is zero on all high-order switching surfaces. On the other hand, the method has several advantages, being faster than the others, and just as accurate.

Its greatest advantage, however, lies in its ability to drive a plant optimally to the origin in face of extremal disturbances. These disturbances are fed into the computer as the variable $w(t)$, which we earlier assumed to be constant. The control law takes such constant disturbances into account when computing the optimal trajectories. Following are graphical comparisons of the time needed to reach the origin disturbance-optimally, with the time needed when a disturbed system is controlled by the no-disturbance law.

Chatter

It will readily be seen that the greater the magnitude of the disturbance, the greater the advantage of disturbance-optimal control. These graphs were prepared by simulating the optimal system within a .001 second time increment, small enough so that finer increments would have no discernible effect on the graph. The non-optimal curve was prepared with a larger increment, .005, because it was discovered that a finer increment actually increased the time to the origin. If the optimal path is as shown in Figure 11 with the net effect of the disturbance w as shown, the non-optimal control will cause the plant to chatter back and forth across the switching surface as shown in Figure 12.

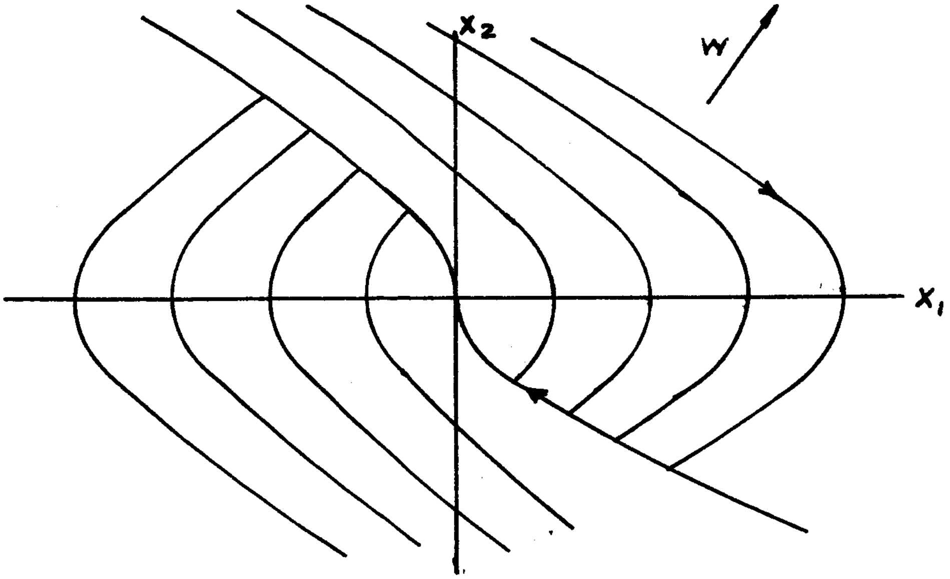


FIGURE 11

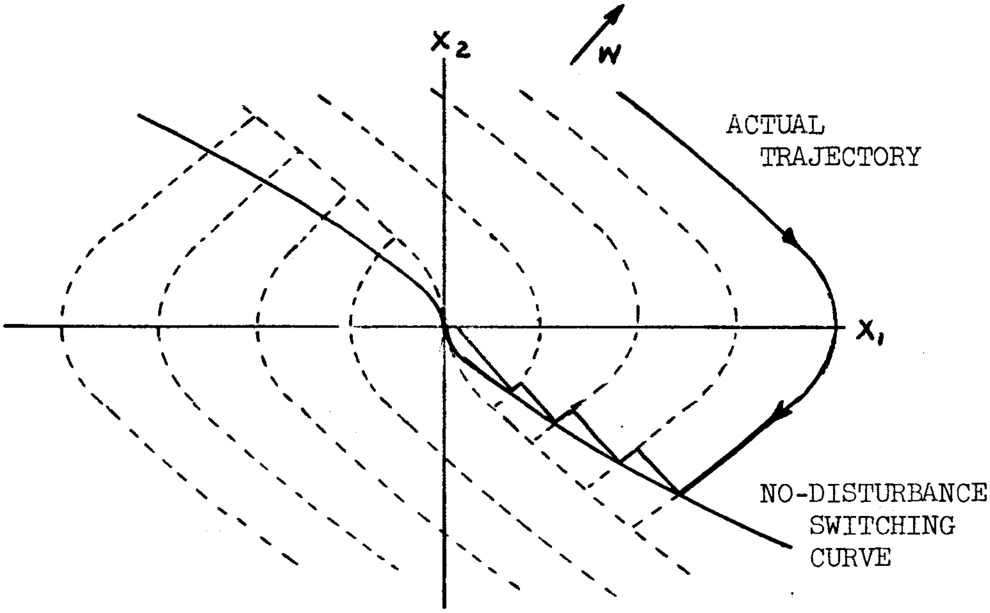


FIGURE 12

The smaller the increment, the more the trajectory will chatter, resulting in a longer time needed to reach the origin. With a .005 increment, up to 17 switches in under .7 seconds were needed to reach the origin.

Sensing Disturbances

The disturbance-optimal control law would be useless without some scheme for actually sensing disturbances as they occur. As a plant moves through phase space, its path is disturbed from the proper path according to the differential equation. It moves normally according to

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-As}b\,\epsilon\,ds$$

But when disturbed, its path is described by

$$x(t) = e^{At}x(0) + e^{At}\int_0^t e^{-As}b(\epsilon + w)ds \tag{31}$$

From (31) we get

$$x_1(t) = x_1(0) + (\epsilon + w)t \tag{32}$$

$$w = \frac{x_1(t)-x_1(0)}{t} - \epsilon \tag{33}$$

This value w is the average disturbance over time t, the constant disturbance which would have moved x(0) to x(t). The disturbance at time t = 0 is given by

$$w = \dot{x}_1(0) - \epsilon \qquad (34)$$

Since this disturbance may be sensed as soon as it occurs, merely by comparing the actual position of the plant to its predicted position, it is feasible to take this disturbance into account when computing optimal trajectories.

Using Disturbance Information

Knowledge of the disturbance level at any time is again useless unless accompanied by some assumption as to its future activity. In the cases tested, it was assumed that the disturbance would remain constant until the plant had been forced into the origin. A slightly better scheme involves taking a new reading of the disturbance at each increment, but still treating each new reading as though w would remain constant at that level until the origin was reaches. However, wind charts compiled at rocket launching sites (since wind is the major disturbance encountered) indicate that the winds most commonly found are like the ones shown in Figures 13 and 14.

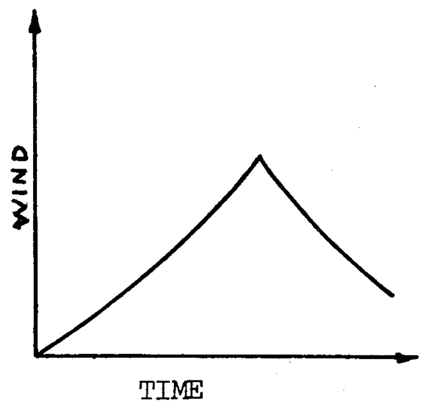


FIGURE 13

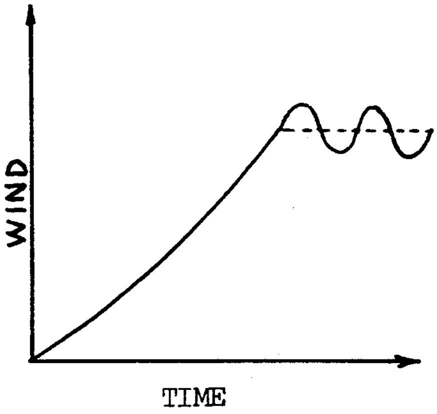


FIGURE 14

The most reasonable approach to the problem of optimiing response with respect to wind would seem to be combining knowledge of the wind chart with on-line sensing. In other words, the best possibility lies in sensing the wind as the plant moves, and from this date, plus knowledge of habitual wind patterns, form good guesses as to the activity of the wind - in the very near future. This is a possible area of future study.

Computer Simulation

The optimal and non-optimal responses with respect to disturbances were run on an IBM 1620 Computer. The computer program for the optimal control law is in Appendix A.

IV. COMPUTER SIMULATION OF THE FOURTH ORDER SYSTEM

As of now, no control law has been formulated for the system

$$x^{(4)} - \lambda^2 \ddot{x} = \epsilon \qquad (35)$$

whose eigenvalues are $0, 0, \lambda, -\lambda$. The numerical methods which were applied to the system

$$\dddot{x} - \lambda^2 \dot{x} = \epsilon, \qquad (36)$$

namely, solving explicitly the equations representing the solution IB and II will not work on this much more complicated system. Since optimal control is not presently available for the $\{0, 0, \lambda, -\lambda\}$ case, it was decided to apply to it non-optimal control, to determine how that compared to the optimal.

The most likely scheme for controlling this system non-optimally seemed to be applying the control law for the third order case (equation 36), whose eigenvalues are $\{0, \lambda, -\lambda\}$ for the third order case, written vectorally as

$$\bar{x}(t) = \overline{Ax} + b\epsilon \qquad (37)$$

where

$$\bar{x} = \begin{vmatrix} x \\ \dot{x} \\ \ddot{x} \end{vmatrix}, \qquad \bar{A} = \begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & \lambda^2 & 0 \end{vmatrix}$$

and $\bar{b} = \begin{vmatrix} 0 \\ 0 \\ 1 \end{vmatrix}$, the control law was found by first transforming $\bar{A}$ to a diagonal form A,

$$A = \begin{vmatrix} 0 & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & -\lambda \end{vmatrix}$$

and $\bar{b}$ to

$$b = \begin{vmatrix} 1 \\ 1 \\ 1 \end{vmatrix}$$

by a matrix Q such that

$$Q^{-1}\bar{A}Q = A$$

and $Q^{-1}\bar{b} = b$.

The matrix Q^{-1} was found to be

$$Q^{-1} = \begin{vmatrix} -1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{vmatrix}$$

The control law operates on the new variables $x' = Q^{-1}x$. However, it was not possible to diagonalize the matrix of the fourth order system, since it had two nondistinct eigenvalues. Therefore, the running proceedure for this test was to move a plant according to the (untransformed) fourth order system, and control the (transformed) variables x_1', x_2', x_3' .

The solution to the fourth order system was as follows:

$x^{(4)} - \lambda^2 \ddot{x} = \epsilon$. Let $\lambda = 1$

Then $x(t) = e^{At}x(0) = e^{At} \int_0^t e^{-As} b \, \epsilon \, ds$

where $e^{At} = \begin{vmatrix} 1 & t & -1 + \cosh t & -t + \sinh t \\ 0 & 1 & \sinh t & -1 + \cosh t \\ 0 & 0 & \cosh t & \sinh t \\ 0 & 0 & \sinh t & \cosh t \end{vmatrix}$

$$e^{At}b = \begin{vmatrix} -t + \sinh t \\ -1 + \cosh t \\ \sinh t \\ \cosh t \end{vmatrix}$$

Let $\xi = x(0) = \int_0^t e^{-As} b\, \epsilon\, ds.$

For an interval $[0,t]$ over which ϵ does not change sign, we have

$$\xi_1 = x_1(0) - \frac{1}{2} t^2 + \cosh t - 1$$

$$\xi_2 = x_2(0) - t + \sinh t$$

$$\xi_3 = x_3(0) + \cosh t - 1$$

$$\xi_4 = x_4(0) + \sinh t$$

Thus $x(t) = e^{At}\xi(t)$ the control law is $\epsilon = \mathrm{sgn}(\sigma)$, where $\sigma = \sigma(x_1', x_2', x_3')$, $\sigma = \sigma(Q^{-1}x)$. The actual formulation of σ is described in IB. To find test points for the fourth order systems, a program was used which found $x(0)$ according to the equations

$$x(t) = e^{At}(x(0) + \int_0^t e^{-As} b\, \epsilon\, ds)$$

When $x(t) = 0$

$$x(0) = - \int_0^t e^{-As} b \, \epsilon \, ds.$$

One could find $x(0)$ from the specifications of the switching times t_1, t_2, t_3, t_4.

At this time, the computer simulation has not worked. One possible reason for the failure is that the control law for the system

$$\dddot{x} - \lambda^2 \dot{x} = \epsilon,$$

which is Laplace-equivalent to

$$\frac{1}{s(s^2-\lambda^2)}$$

might better be used to control

$$x^{(4)} - \lambda^2 \dot{x} = \dot{\epsilon} + \alpha\epsilon, \quad \alpha < k < 1,$$

or

$$\frac{s + \alpha}{s^2(s^2-\lambda^2)}$$

This is because for small α,

$$\frac{s + \alpha}{s(s^2 - \lambda^2)} \doteq \frac{1}{s(s^2 - \lambda^2)}$$

This new control scheme has not yet been investigated.

APPENDIX A

```
C  CONTROL LAW III WITH STEP NOISE
1  FORMAT (F10.4,F10.4,F10.4,F10.4)
2  FORMAT (6X 14,F10.2,F10.4/)
30 FORMAT (6H MIN, F10.4,F10.4/)
31 FORMAT (F10.4)
   DIMENSION T(20)
3  ACCEPT 1, XX1,XX2,XX3,DT
33 ACCEPT 31,EPM
   EPN=EPM
32 X1=XX1
   X2=XX2
   X3=XX3
   TYPE 31,EPN
   DO 4 I 1,20
4  T(I)= 0.
   K=1
   D=1.
   EP1=1.
   SIG1 = 1.
   D1=10000.
   TT=0.
   I=1
   GO TO 21
6  IF(X1)5,7,5
5  EP1=-SIGN(X1)
7  IF(SIGN(EP1-EXP(-EP1*X1)*(X2 + EP1)))8,10,9
8  SIG1=-1.
   GO TO 10
9  SIG1=1.
10 Z=EXP(SIG1*X1)
   SIG2=SIG1 + 1,/Z*(X2 + SIG1)
   SIG3=SIG1 + Z*(X3-SIG1)
   Z=1. + SIG2*SIG1
   SIGMA=-((SIG3*Z-SIG2)**2 + 4.*SIG2*SIG3)*SIG1
   SIGMA=SIGMA*(SIG3*Z**2*SIG1-SIG2**2+ SIG1*SIG2 + 2.*ABS(SIG2)**1.5)
   GO TO (11,12),K
```

```
11 EPSI=SIGN(SIGMA) + EPN
   K=2
12 D=X1**2 + X2**2 + X3**2
   IF(SIGMA*EPSI)20,13,13
13 IF(I-2)17,17,14
14 IF(D1-D)*D2)15,15,24
15 DR=SQRT(D1)
   TR=TT-DT
   TYPE30,I,TR,DR
   D2=-D2
   IF(SENSE SWITCH 1)23,24
24 D1=D
17 GO TO (18,19),J
18 C2=X20 + EPSI
   C3=EPSI-X30
   J=2
19 TI=T(I) + DT
   TT=TT + DT
   E=EXP(T(I))
   XI=EPSI*T(I) + X10
   X2=C2*E-EPSI
   X3=EPSI-C3/E
   GO TO 6
20 EPSI=SIGN(SIGMA) + EPN
   DR=SQRT(D)
   TYPE2,I,TT,DR
   I=I + 1
   D2=1
21 X10=X1
   X20=X2
   X30=X3
   J=1
   GO TO 6
23 (IF(SENSE SWITCH 3)33,16
16 EPN=EPN + EPM
   IF(SENSE SWITCH 2)3,32
   END
```

```
C  GILCHRIST CONTROL LAW III
1  FORMAT (F10.4,F10.4,F10.4,F10.4,F10.4)
2  FORMAT (F10.4,F10.4,F10.4)
3  FORMAT (F10.4,F10.4,E14.4/)
   DIMENSION T(3),E(3)
5  ACCEPT 1,X1,X2,X3,DT
6  ACCEPT 2,A,B,C
   D=3./A
   E(2)=EXP(-D-A)
   E(3)=EXP(-A + D)
   E(1)=EXP(-A*D)
   RF=C/A*(E(1)-1.) + X1
   RS=B + C/(1.+ A)*E(2)-1.) + X2
   RT=-B-C/(1.-A)*(E(3)-1.) + X3
   U=1.
7  R1=U*RF
   R2=U*RS-1.
   R3=U*RT-1.
   F=1.+ B*U
   TYPE 2,R1,R2,R3
   S1=-1.
   T(3)=0.
9  E(3)=EXP(T(3))
   E3=EXP(.5*(F*T(3) + R1))
   ER1=2.*(E3-1.)/(F*E(3) + R3)
   ER2=ER1/E3
   G=2.*(ER1-ER2) + F/E(3) + R2
   IF(S1)10,12,15
10 S1=0.
   G1=G
   TYPE 3,T(3),ER1,G
11 T(3)=T(3) + DT
   IF(D-T(3))9,9,11
12 S1=1.
   TYPE 3,T(3),ER1,G
   IF(G1*G)13,14,14
13 PAUSE
14 G1=G
15 IF(SENSE SWITCH 1)16,18
16 TYPE 3,T(3),ER1,G
   IF(SENSE SWITCH 2)17,18
17 U=-U
   GO TO 7
```

```
18 IF(G1*G)20,20,19
19 T(3)=T(3) + DT
   G1=G
   GO TO 9
20 TYPE 3,T(3),ER1,G
   E(1)=1./ER1
   E(2)=1./ER2
   T(1)=LOG(E(1))
   T(2)=LOG(E(2))
   R1=U*(2.*(T(1)-T(2)) + F*T(3) + R1)
   R2=E(3)*G
   R(3)=(2.*(E(1)-E(2)) + R3)/E(3) + F
   D=SQRT(R1**2 + R2**2 + R3**2)
   TYPE 1,T(1),T(2),T(3),D
   IF(SENSE SWITCH 2)5,6
   END
```

WIND OPTIMAL SEEING EYE CONTROL WITH HELP

```
1  FORMAT (F10.4,F10.4,F10.4,F10.4)
2  FORMAT (F10.4)
3  FORMAT (6X I4,F10.4,F10.4/)
4  FORMAT (6H MIN,I4,F10.4,F10.4/)
30 FORMAT (I4)
   DIMENSION TA(20)
   DO 40 I=5,20
40 TA(I) = 0.
5  ACCEPT 1,XX1,XX2,XX3,DT
   ACCEPT 30,K
   ET=EXP(DT)
   ET1=ET-1.
   ERT=1./ET
   ERT1=ERT-1.
6  ACCEPT 2,U
   ACCEPT 2,W
   TT=0.
   I=1
   D1=100.
   X1=XX1
   X2=XX2
   X3=XX3
   ACCEPT 1,TA(1),TA(2),TA(3),TA(4)
   GO TO 32
8  S=-1.
   S1=-1.
   D2=1.
   T1=0.
9  IF(I-K)19,19,14
14 D=X1**2 + X2**2 + X3**2
   IF(D2*(D1-D))15,15,18
15 IF(D2)17,17,16
16 A=SQRT(D1)
   T2=TT-DT
   TYPE 4,I,T2,A
   IF(SENSE SWITCH 2)26,17
17 D2=-D2
18 D1=D
19 R1=-U*X1
   R2=U*(X2 + W)-1.
```

```
    R3=-U*(X3-W)-1.
    G=1.+ U*W
    A=2.*R2
    B=4.-G**2 + R2*R3
    DSC=B**2-8.*A*R3
    IF(DSC)10,11,11
 10 IF(S)101,101,22
101 S=1.
    IF(S1)102,102,22
102 F1=-1.
    S1=1.
    GO TO 22
 11 E2=.5*(-B-SQRT(DSC))/A
    E3=(2.*E2 + R3)/G
    T2=LOG(E2)
    T3=LOG(E3)
    F=2.*T2-G*T3 + R1
    IF(S1)12,13,13
 12 F1=SIGN(F)
    F1=SIGN(F1 + .5)
    S1=1.
 13 IF(F*F1)25,20,20
 20 IF(SENSE SWITCH 1)21,22
 21 TYPE 3,I,TT,F
 22 T1=T1 + DT
    TT=TT + DT
    IF(SENSE SWITCH 3)23,24
 23 TYPE 2,TT
    ACCEPT 2,W
 24 A=U + W
    X1=X1 + DT*A
    X2=ET*(X2-A*ERT1)
    X3=ERT*(X3 + A*ET1)
    GO TO 9
 25 IF(T3-T2)31,28,28
 28 IF(T2)31,29,29
 29 TYPE 3,I,TT,F
    I=I + 1
    U=-U
```

```
32 A=U + W
   R1=EXP(TA(I))
   R2=1./R1
   X1=X1 + A*TA(I)
   X2=R1*(X2-A*(R2-1.))
   X3=R2*(X3 + A*(R1-1.))
   GO TO 8
26 IF(SENSE SWITCH 3)5,6
31 F1=-F1
   GO TO 22
   END
```